International Journal on Advances in Intelligent Systems

2015 vol. 8 nr. 1&2

The *International Journal on Advances in Intelligent Systems* is Published by IARIA.
ISSN: 1942-2679
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Intelligent Systems
Volume 8, Number 1 & 2, 2015

Ateet Bhalla, Oriental Institute of Science & Technology, Bhopal, India
Freimut Bodendorf, Universität Erlangen-Nürnberg, Germany
Karsten Böhm, FH Kufstein Tirol - University of Applied Sciences, Austria
Pierre Borne, Ecole Centrale de Lille, France
Christos Bouras, University of Patras, Greece
Anne Boyer, LORIA - Nancy Université / KIWI Research team, France
Stainam Brandao, COPPE/Federal University of Rio de Janeiro, Brazil
Stefano Bromuri, University of Applied Sciences Western Switzerland, Switzerland
Vít Bršlica, University of Defence - Brno, Czech Republic
Dumitru Burdescu, University of Craiova, Romania
Diletta Romana Cacciagrano, University of Camerino, Italy
Kenneth P. Camilleri, University of Malta - Msida, Malta
Paolo Campegiani, University of Rome Tor Vergata , Italy
Marcelino Campos Oliveira Silva, Chemtech - A Siemens Business / Federal University of Rio de Janeiro, Brazil
Ozgu Can, Ege University, Turkey
José Manuel Cantera Fonseca, Telefónica Investigación y Desarrollo (R&D), Spain
Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Miriam A. M. Capretz, The University of Western Ontario, Canada
Massimiliano Caramia, University of Rome "Tor Vergata", Italy
Davide Carboni, CRS4 Research Center - Sardinia, Italy
Luis Carriço, University of Lisbon, Portugal
Rafael Casado Gonzalez, Universidad de Castilla - La Mancha, Spain
Michelangelo Ceci, University of Bari, Italy
Fernando Cerdan, Polytechnic University of Cartagena, Spain
Alexandra Suzana Cernian, University "Politehnica" of Bucharest, Romania
Sukalpa Chanda, Gjøvik University College, Norway
David Chen, University Bordeaux 1, France
Po-Hsun Cheng, National Kaohsiung Normal University, Taiwan
Dickson Chiu, Dickson Computer Systems, Hong Kong
Sunil Choenni, Research & Documentation Centre, Ministry of Security and Justice / Rotterdam University of Applied Sciences, The Netherlands
Ryszard S. Choras, University of Technology & Life Sciences, Poland
Smitashree Choudhury, Knowledge Media Institute, The UK Open University, UK
William Cheng-Chung Chu, Tunghai University, Taiwan
Christophe Claramunt, Naval Academy Research Institute, France
Cesar A. Collazos, Universidad del Cauca, Colombia
Phan Cong-Vinh, NTT University, Vietnam
Christophe Cruz, University of Bourgogne, France
Beata Czarnacka-Chrobot, Warsaw School of Economics, Department of Business Informatics, Poland
Claudia d'Amato, University of Bari, Italy
Sérgio Roberto P. da Silva, Universidade Estadual de Maringá - Paraná, Brazil
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Dragos Datcu, Netherlands Defense Academy / Delft University of Technology , The Netherlands
Antonio De Nicola, ENEA, Italy
Claudio de Castro Monteiro, Federal Institute of Education, Science and Technology of Tocantins, Brazil
Noel De Palma, Joseph Fourier University, France

Zhi-Hong Deng, Peking University, China
Stojan Denic, Toshiba Research Europe Limited, UK
Vivek S. Deshpande, MIT College of Engineering - Pune, India
Sotirios Ch. Diamantas, Pusan National University, South Korea
Leandro Dias da Silva, Universidade Federal de Alagoas, Brazil
Jerome Dinet, Univeristé Paul Verlaine - Metz, France
Jianguo Ding, University of Luxembourg, Luxembourg
Yulin Ding, Defence Science & Technology Organisation Edinburgh, Australia
Alexiei Dingli, University of Malta, Malta
Mihaela Dinsoreanu, Technical University of Cluj-Napoca, Romania
Ioanna Dionysiou, University of Nicosia, Cyprus
Roland Dodd, CQUniversity, Australia
Nima Dokoohaki, Royal Institute of Technology (KTH)-Kista, Sweden
Suzana Dragicevic, Simon Fraser University- Burnaby, Canada
Mauro Dragone, University College Dublin (UCD), Ireland
Marek J. Druzdzel, University of Pittsburgh, USA
Carlos Duarte, University of Lisbon, Portugal
Raimund K. Ege, Northern Illinois University, USA
Jorge Ejarque, Barcelona Supercomputing Center, Spain
Larbi Esmahi, Athabasca University, Canada
Simon G. Fabri, University of Malta, Malta
Umar Farooq, Amazon.com, USA
Mehdi Farshbaf-Sahih-Sorkhabi, Azad University - Tehran / Fanavaran co., Tehran, Iran
Anna Fensel, Semantic Technology Institute (STI) Innsbruck and FTW Forschungszentrum Telekommunikation Wien, Austria
Stenio Fernandes, Federal University of Pernambuco (CIn/UFPE), Brazil
Oscar Ferrandez Escamez, University of Utah, USA
Agata Filipowska, Poznan University of Economics, Poland
Ziny Flikop, Scientist, USA
Adina Magda Florea, University "Politehnica" of Bucharest, Romania
Francesco Fontanella, University of Cassino and Southern Lazio, Italy
Panagiotis Fotaris, University of Macedonia, Greece
Enrico Francesconi, ITTIG - CNR / Institute of Legal Information Theory and Techniques / Italian National Research Council, Italy
Rita Francese, Università di Salerno - Fisciano, Italy
Bernhard Freudenthaler, Software Competence Center Hagenberg GmbH, Austria
Sören Frey, Daimler TSS GmbH, Germany
Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Somchart Fugkeaw, Thai Digital ID Co., Ltd., Thailand
Naoki Fukuta, Shizuoka University, Japan
Mathias Funk, Eindhoven University of Technology, The Netherlands
Adam M. Gadomski, Università degli Studi di Roma La Sapienza, Italy
Alex Galis, University College London (UCL), UK
Crescenzio Gallo, Department of Clinical and Experimental Medicine - University of Foggia, Italy
Matjaz Gams, Jozef Stefan Institute-Ljubljana, Slovenia
Raúl García Castro, Universidad Politécnica de Madrid, Spain

Fabio Gasparetti, Roma Tre University - Artificial Intelligence Lab, Italy
Joseph A. Giampapa, Carnegie Mellon University, USA
George Giannakopoulos, NCSR Demokritos, Greece
David Gil, University of Alicante, Spain
Harald Gjermundrod, University of Nicosia, Cyprus
Angelantonio Gnazzo, Telecom Italia - Torino, Italy
Luis Gomes, Universidade Nova Lisboa, Portugal
Nan-Wei Gong, MIT Media Laboratory, USA
Francisco Alejandro Gonzale-Horta, National Institute for Astrophysics, Optics, and Electronics (INAOE), Mexico
Sotirios K. Goudos, Aristotle University of Thessaloniki, Greece
Victor Govindaswamy, Texas A&M University-Texarkana, USA
Gregor Grambow, University of Ulm, Germany
Fabio Grandi, University of Bologna, Italy
Andrina Granić, University of Split, Croatia
Carmine Gravino, Università degli Studi di Salerno, Italy
Michael Grottke, University of Erlangen-Nuremberg, Germany
Vic Grout, Glyndŵr University, UK
Maik Günther, Stadtwerke München GmbH, Germany
Francesco Guerra, University of Modena and Reggio Emilia, Italy
Alessio Gugliotta, Innova SPA, Italy
Richard Gunstone, Bournemouth University, UK
Fikret Gurgen, Bogazici University, Turkey
Maki Habib, The American University in Cairo, Egypt
Till Halbach Røssvoll, Norwegian Computing Center, Norway
Jameleddine Hassine, King Fahd University of Petroleum & Mineral (KFUPM), Saudi Arabia
Ourania Hatzi, Harokopio University of Athens, Greece
Yulan He, Aston University, UK
Kari Heikkinen, Lappeenranta University of Technology, Finland
Cory Henson, Wright State University / Kno.e.sis Center, USA
Arthur Herzog, Technische Universität Darmstadt, Germany
Rattikorn Hewett, Whitacre College of Engineering, Texas Tech University, USA
Celso Massaki Hirata, Instituto Tecnológico de Aeronáutica - São José dos Campos, Brazil
Jochen Hirth, University of Kaiserslautern, Germany
Bernhard Hollunder, Hochschule Furtwangen University, Germany
Thomas Holz, University College Dublin, Ireland
Władysław Homenda, Warsaw University of Technology, Poland
Carolina Howard Felicíssimo, Schlumberger Brazil Research and Geoengineering Center, Brazil
Jingwei Huang , University of Illinois at Urbana-Champaign, USA
Weidong (Tony) Huang, CSIRO ICT Centre, Australia
Xiaodi Huang, Charles Sturt University - Albury, Australia
Eduardo Huedo, Universidad Complutense de Madrid, Spain
Marc-Philippe Huget, University of Savoie, France
Chi Hung, Tsinghua University, China
Chih-Cheng Hung, Southern Polytechnic State University - Marietta, USA
Edward Hung, Hong Kong Polytechnic University, Hong Kong
Muhammad Iftikhar, Universiti Malaysia Sabah (UMS), Malaysia

Prateek Jain, Ohio Center of Excellence in Knowledge-enabled Computing, Kno.e.sis, USA
Wassim Jaziri, Miracl Laboratory, ISIM Sfax, Tunisia
Hoyoung Jeung, SAP Research Brisbane, Australia
Yiming Ji, University of South Carolina Beaufort, USA
Jinlei Jiang, Department of Computer Science and Technology, Tsinghua University, China
Weirong Jiang, Juniper Networks Inc., USA
Hanmin Jung, Korea Institute of Science & Technology Information, Korea
Ilya S. Kabak, "Stankin" Moscow State Technological University, Russia
Eleanna Kafeza, Athens University of Economics and Business, Greece
Hermann Kaindl, Vienna University of Technology, Austria
Ahmed Kamel, Concordia College, Moorhead, Minnesota, USA
Rajkumar Kannan, Bishop Heber College(Autonomous), India
Fazal Wahab Karam, Norwegian University of Science and Technology (NTNU), Norway
Dimitrios A. Karras, Chalkis Institute of Technology, Hellas
Koji Kashihara, The University of Tokushima, Japan
Nittaya Kerdprasop, Suranaree University of Technology, Thailand
Katia Kermanidis, Ionian University, Greece
Serge Kernbach, University of Stuttgart, Germany
Nhien An Le Khac, University College Dublin, Ireland
Reinhard Klemm, Avaya Labs Research, USA
Ah-Lian Kor, Leeds Metropolitan University, UK
Arne Koschel, Applied University of Sciences and Arts, Hannover, Germany
George Kousiouris, NTUA, Greece
Philipp Kremer, German Aerospace Center (DLR), Germany
Dalia Kriksciuniene, Vilnius University, Lithuania
Dariusz Król, AGH University of Science and Technology, ACC Cyfronet AGH, Poland
Markus Kunde, German Aerospace Center, Germany
Dharmender Singh Kushwaha, Motilal Nehru National Institute of Technology, India
Andrew Kusiak, The University of Iowa, USA
Dimosthenis Kyriazis, National Technical University of Athens, Greece
Vitaveska Lanfranchi, Research Fellow, OAK Group, University of Sheffield, UK
Mikel Larrea, University of the Basque Country UPV/EHU, Spain
Philippe Le Parc, University of Brest, France
Gyu Myoung Lee, Liverpool John Moores University, UK
Kyu-Chul Lee, Chungnam National University, South Korea
Tracey Kah Mein Lee, Singapore Polytechnic, Republic of Singapore
Daniel Lemire, LICEF Research Center, Canada
Haim Levkowitz, University of Massachusetts Lowell, USA
Kuan-Ching Li, Providence University, Taiwan
Tsai-Yen Li, National Chengchi University, Taiwan
Yangmin Li, University of Macau, Macao SAR
Jian Liang, Nimbus Centre, Cork Institute of Technology, Ireland
Haibin Liu, China Aerospace Science and Technology Corporation, China
Lu Liu, University of Derby, UK
Qing Liu, The Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Shih-Hsi "Alex" Liu, California State University - Fresno, USA

Xiaoqing (Frank) Liu, Missouri University of Science and Technology, USA
David Lizcano, Universidad a Distancia de Madrid, Spain
Henrique Lopes Cardoso, LIACC / Faculty of Engineering, University of Porto, Portugal
Sandra Lovrencic, University of Zagreb, Croatia
Jun Luo, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
Prabhat K. Mahanti, University of New Brunswick, Canada
Jacek Mandziuk, Warsaw University of Technology, Poland
Herwig Mannaert, University of Antwerp, Belgium
Yannis Manolopoulos, Aristotle University of Thessaloniki, Greece
Antonio Maria Rinaldi, Università di Napoli Federico II, Italy
Ali Masoudi-Nejad, University of Tehran, Iran
Constandinos Mavromoustakis, University of Nicosia, Cyprus
Zulfiqar Ali Memon, Sukkur Institute of Business Administration, Pakistan
Andreas Merentitis, AGT Group (R&D) GmbH, Germany
Jose Merseguer, Universidad de Zaragoza, Spain
Frederic Migeon, IRIT/Toulouse University, France
Harald Milchrahm, Technical University Graz, Institute for Software Technology, Austria
Les Miller, Iowa State University, USA
Marius Minea, University POLITEHNICA of Bucharest, Romania
Yasser F. O. Mohammad, Assiut University, Egypt
Shahab Mokarizadeh, Royal Institute of Technology (KTH) - Stockholm, Sweden
Martin Molhanec, Czech Technical University in Prague, Czech Republic
Charalampos Moschopoulos, KU Leuven, Belgium
Mary Luz Mouronte López, Ericsson S.A., Spain
Henning Müller, University of Applied Sciences Western Switzerland - Sierre (HES SO), Switzerland
Susana Munoz Hernández, Universidad Politécnica de Madrid, Spain
Adrian Muscat, University of Malta, Malta
Peter Mutschke, GESIS - Leibniz Institute for the Social Sciences - Bonn, Germany
Bela Mutschler, Hochschule Ravensburg-Weingarten, Germany
Deok Hee Nam, Wilberforce University, USA
Fazel Naghdy, University of Wollongong, Australia
Joan Navarro, Research Group in Distributed Systems (La Salle - Ramon Llull University), Spain
Rui Neves Madeira, Instituto Politécnico de Setúbal / Universidade Nova de Lisboa, Portugal
Toàn Nguyên, INRIA Grenobel Rhone-Alpes/ Montbonnot, France
Andrzej Niesler, Institute of Business Informatics, Wroclaw University of Economics, Poland
Kouzou Ohara, Aoyama Gakuin University, Japan
Jonice Oliveira, Universidade Federal do Rio de Janeiro, Brazil
Ian Oliver, Nokia Location & Commerce, Finland / University of Brighton, UK
Michael Adeyeye Oluwasegun, University of Cape Town, South Africa
Sigeru Omatu, Osaka Institute of Technology, Japan
Sascha Opletal, University of Stuttgart, Germany
Fakri Othman, Cardiff Metropolitan University, UK
Enn Õunapuu, Tallinn University of Technology, Estonia
Jeffrey Junfeng Pan, Facebook Inc., USA
Hervé Panetto, University of Lorraine, France
Malgorzata Pankowska, University of Economics, Poland

Harris Papadopoulos, Frederick University, Cyprus
Laura Papaleo, ICT Department - Province of Genoa & University of Genoa, Italy
Agis Papantoniou, National Technical University of Athens, Greece
Thanasis G. Papaioannou, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Andreas Papasalouros, University of the Aegean, Greece
Eric Paquet, National Research Council / University of Ottawa, Canada
Kunal Patel, Ingenuity Systems, USA
Carlos Pedrinaci, Knowledge Media Institute, The Open University, UK
Yoseba Penya, University of Deusto - DeustoTech (Basque Country), Spain
Cathryn Peoples, University of Ulster, UK
Asier Perallos, University of Deusto, Spain
Christian Percebois, Université Paul Sabatier - IRIT, France
Andrea Perego, European Commission, Joint Research Centre, Italy
Mark Perry, University of Western Ontario/Faculty of Law/ Faculty of Science - London, Canada
Willy Picard, Poznań University of Economics, Poland
Meikel Poess, Oracle, USA
Agostino Poggi, Università degli Studi di Parma, Italy
R. Ponnusamy, Madha Engineering College-Anna University, India
Dorin Popescu, University of Craiova, Romania
Stefan Poslad, Queen Mary University of London, UK
Wendy Powley, Queen's University, Canada
Radu-Emil Precup, "Politehnica" University of Timisoara, Romania
Jerzy Prekurat, Canadian Bank Note Co. Ltd., Canada
Didier Puzenat, Université des Antilles et de la Guyane, France
Sita Ramakrishnan, Monash University, Australia
Elmano Ramalho Cavalcanti, Federal University of Campina Grande, Brazil
Juwel Rana, Luleå University of Technology, Sweden
Martin Randles, School of Computing and Mathematical Sciences, Liverpool John Moores University, UK
Christoph Rasche, University of Paderborn, Germany
Ann Reddipogu, ManyWorlds UK Ltd, UK
Ramana Reddy, West Virginia University, USA
René Reiners, Fraunhofer FIT - Sankt Augustin, Germany
Paolo Remagnino, Kingston University - Surrey, UK
Sebastian Rieger, University of Applied Sciences Fulda, Germany
Andreas Riener, Johannes Kepler University Linz, Austria
Ivan Rodero, NSF Center for Autonomic Computing, Rutgers University - Piscataway, USA
Joel Rodrigues, Instituto de Telecomunicações / University of Beira Interior, Portugal
Alejandro Rodríguez González, University Carlos III of Madrid, Spain
Paolo Romano, INESC-ID Lisbon, Portugal
Agostinho Rosa, Instituto de Sistemas e Robótica, Portugal
José Rouillard, University of Lille, France
Paweł Różycki, University of Information Technology and Management (UITM) in Rzeszów, Poland
Igor Ruiz-Agundez, DeustoTech, University of Deusto, Spain
Michele Ruta, Politecnico di Bari, Italy
Melike Sah, Trinity College Dublin, Ireland
Francesc Saigi Rubió, Universitat Oberta de Catalunya, Spain

Abdel-Badeeh M. Salem, Ain Shams University, Egypt
Yacine Sam, Université François-Rabelais Tours, France
Ismael Sanz, Universitat Jaume I, Spain
Ricardo Sanz, Universidad Politecnica de Madrid, Spain
Marcello Sarini, Università degli Studi Milano-Bicocca - Milano, Italy
Munehiko Sasajima, I.S.I.R., Osaka University, Japan
Minoru Sasaki, Ibaraki University, Japan
Hiroyuki Sato, University of Tokyo, Japan
Jürgen Sauer, Universität Oldenburg, Germany
Patrick Sayd, CEA List, France
Dominique Scapin, INRIA - Le Chesnay, France
Kenneth Scerri, University of Malta, Malta
Adriana Schiopoiu Burlea, University of Craiova, Romania
Rainer Schmidt, Austrian Institute of Technology, Austria
Bruno Schulze, National Laboratory for Scientific Computing - LNCC, Brazil
Wieland Schwinger, Johannes Kepler University Linz, Austria
Hans-Werner Sehring, T-Systems Multimedia Solutions GmbH, Germany
Paulo Jorge Sequeira Gonçalves, Polytechnic Institute of Castelo Branco, Portugal
Sandra Sendra Compte, Polytechnic University of Valencia, Spain
Kewei Sha, Oklahoma City University, USA
Roman Y. Shtykh, Rakuten, Inc., Japan
Robin JS Sloan, University of Abertay Dundee, UK
Vasco N. G. J. Soares, Instituto de Telecomunicações / University of Beira Interior / Polytechnic Institute of Castelo Branco, Portugal
Don Sofge, Naval Research Laboratory, USA
Christoph Sondermann-Woelke, Universitaet Paderborn, Germany
George Spanoudakis, City University London, UK
Vladimir Stantchev, SRH University Berlin, Germany
Claudius Stern, University of Paderborn, Germany
Mari Carmen Suárez-Figueroa, Universidad Politécnica de Madrid (UPM), Spain
Kåre Synnes, Luleå University of Technology, Sweden
Ryszard Tadeusiewicz, AGH University of Science and Technology, Poland
Yehia Taher, ERISS - Tilburg University, The Netherlands
Yutaka Takahashi, Senshu University, Japan
Dan Tamir, Texas State University, USA
Jinhui Tang, Nanjing University of Science and Technology, P.R. China
Yi Tang, Chinese Academy of Sciences, China
John Terzakis, Intel, USA
Sotirios Terzis, University of Strathclyde, UK
Vagan Terziyan, University of Jyvaskyla, Finland
Ioan Toma, STI Innsbruck/University Innsbruck, Austria
Lucio Tommaso De Paolis, Department of Innovation Engineering - University of Salento, Italy
Davide Tosi, Università degli Studi dell'Insubria, Italy
Raquel Trillo Lado, University of Zaragoza, Spain
Tuan Anh Trinh, Budapest University of Technology and Economics, Hungary
Simon Tsang, Applied Communication Sciences, USA

Theodore Tsiligiridis, Agricultural University of Athens, Greece
Antonios Tsourdos, Cranfield University, UK
José Valente de Oliveira, University of Algarve, Portugal
Eugen Volk, University of Stuttgart, Germany
Mihaela Vranić, University of Zagreb, Croatia
Chieh-Yih Wan, Intel Labs, Intel Corporation, USA
Jue Wang, Washington University in St. Louis, USA
Shenghui Wang, OCLC Leiden, The Netherlands
Zhonglei Wang, Karlsruhe Institute of Technology (KIT), Germany
Laurent Wendling, University Descartes (Paris 5), France
Maarten Weyn, University of Antwerp, Belgium
Nancy Wiegand, University of Wisconsin-Madison, USA
Alexander Wijesinha, Towson University, USA
Eric B. Wolf, US Geological Survey, Center for Excellence in GIScience, USA
Ouri Wolfson, University of Illinois at Chicago, USA
Yingcai Xiao, The University of Akron, USA
Reuven Yagel, The Jerusalem College of Engineering, Israel
Fan Yang, Nuance Communications, Inc., USA
Maribel Yasmina Santos, University of Minho, Portugal
Zhenzhen Ye, Systems & Technology Group, IBM, US A
Jong P. Yoon, MATH/CIS Dept, Mercy College, USA
Shigang Yue, School of Computer Science, University of Lincoln, UK
Constantin-Bala Zamfirescu, "Lucian Blaga" Univ. of Sibiu, Romania
Claudia Zapata, Pontificia Universidad Católica del Perú, Peru
Marek Zaremba, University of Quebec, Canada
Filip Zavoral, Charles University Prague, Czech Republic
Yuting Zhao, University of Aberdeen, UK
Hai-Tao Zheng, Graduate School at Shenzhen, Tsinghua University, China
Yu Zheng, Microsoft Research Asia, China
Zibin (Ben) Zheng, Shenzhen Research Institute, The Chinese University of Hong Kong, Hong Kong
Bin Zhou, University of Maryland, Baltimore County, USA
Alfred Zimmermann, Reutlingen University - Faculty of Informatics, Germany
Wolf Zimmermann, Martin-Luther-University Halle-Wittenberg, Germany

International Journal on Advances in Intelligent Systems
Volume 8, Numbers 1 & 2, 2015

CONTENTS

A Smart Waste Management with Self-Describing Complex Objects

Yann Glouche, Arnab Sinha, and Paul Couderc
INRIA, Unité de Recherche Rennes-Bretagne-Atlantique
Campus de Beaulieu, Rennes, France
email: {yann.glouche,arnab.sinha,paul.couderc}@inria.fr

***Abstract*—Radio Frequency Identification (RFID) is a pervasive computing technology that can be used to improve waste management by providing early automatic identification of waste at bin level. In this paper, we have presented a smart bin application based on information self-contained in tags associated to each waste item. The wastes are tracked by smart bins using a RFID-based system without requiring the support of an external information system. Two crucial features of the selective sorting process can be improved by using this approach. First, the user is helped in the application of selective sorting. Second, the smart bin knows its content up to the precision of composed materials by types and percentage. It can report back with its status or abnormalities to the rest of the recycling chain. Complex objects like e-waste, hazardous ones, etc. can also be sorted and detected for hazards with the self-describing approach.**

Keywords-green IT; waste management; recycling chain; RFID; NFC; QR code.

I. Introduction

Waste management is an important requirement for ecologically sustainable development of many countries. Efficient sorting of waste is a major issue in today's society. In [1], the concept of self-describing objects is introduced for using the technologies of information and communication to improve the recycling process. In Europe, the consumer society has led to an ever increasing production of waste [2]. This is a consequence of the consumer's behavior, and is worsened by packaging. In [3], it is shown, that the production of waste reaches almost 1.2 kg/day/inhabitant in western Europe. Paradoxically, the same consumers who are concerned with environmental protection are often reluctant when it comes to have more land-filing or more incinerators. Therefore, waste should be disposed and treated properly to reduce environmental impact.

Waste management services are becoming an important market, for which the waste collection process is a critical aspect for the service providers [4], [5]. The main goals are the following :

1) Reducing waste production
2) Ensuring that wastes are properly disposed
3) Recycling and re-using disposed products

To achieve these goals, regulations and taxes are being implemented to favor virtuous behaviors. In particular, to reduce the production of waste, there is an increasing trend towards individual billing, where people are charged depending on waste quantity disposed.

Selective sorting is another approach, which is often implemented to improve recycling and reduce the environment impact. The importance of resources and energy saving is another argument to manufacture recyclable materials.

The sorting of wastes must be implemented as early as possible in the chain to increase the quantity of valuable recyclable materials. The use of pervasive computing technology such as Radio Frequency Identification (RFID) and sensor networks offer a new way to optimize the waste management systems.

In recent years, we have seen increasing adoption of the RFID technology in many application domains, such as logistic, inventory, public transportation and security. Essentially, RFID makes it possible to read digital information from one or several objects using a reader within proximity of the objects, enabling automatic identification, tracking, checking of properties, etc. Apart from this, RFID has added advantages over barcodes. While barcodes compulsorily acts as an enabler that links to retailer's/manufacturer's centralized data (mostly exclusive), RFID can mimic the same with an additional advantage of having a memory for storing some information locally. This locally attached related information could be easily accessed by end-users; an evolution of QR code. Hence, it could be predicted that RFID could replace existing barcodes, QR codes, and attached to most products by the entities for better handling. In this perspective, it is the perfect time to use RFID for waste domain and leverage from their properties to improve current waste management processes.

This paper demonstrates a method to improve the quality of selective sorting. The approach is based on local interactions to track the waste flow of a city. Each waste is detected by information properties stored in a RFID tag associated to it. At each step where wastes are to be processed the RFID tags are read in order to provide the relevant information. This process improves the reuse of recyclable products. We assume that organic waste products are not recycled and hence RFID tags are not attached to them.

One of the advantages of the approach is that it improves the sorting quality without using an external information system. Rather the information is distributed locally in the physical space within the tag memory associated to each waste; thereby increasing the availability of information for various purposes. For example, to help the user in the sorting process and to analyze the content of a bin etc.

This article is organized as follows. The next section outlines the architecture used to process the waste flow in our waste management system. Then, we present a tagged (or "self-describing") waste approach and its use in waste sorting system. The fourth section illustrates a certification process of the content giving reward to the users participating in the waste selective sorting. The fifth section discusses other solutions to sort more complex waste (objects). The sixth section presents the communication system between bins and the recycling service provider with the prototype demonstration illustrated next. Section VIII presents the related work along with a contrast, highlighting the novelty of our work. Finally, Section IX concludes the paper.

II. Waste flow and global architecture of the waste management system

Demonstrating efficient waste management solutions is the primary goal of this article. These solutions are specific to the different phases that pieces of waste undergoes in the system, discussed later in this section. However, all these solutions exploit our principal approach of self-describing objects. Most of these everyday used objects also undergo through different other phases in their life cycle; from manufacturing until disposal. As mentioned earlier, manufacturers and retailers already use RFID tagging of their products extensively for inventory. We have assumed their use would be extended with self-description to make them smarter. This would enable them to participate in smart interactions during their product phase as well after their disposal; when they become waste. Hence, the tangible data for these items would be available pervasively for autonomous processing throughout the waste management chain. Having said this, it should be noted that with the flow of waste across its management chain, their self-describing information (or tangible data) also gets aggregated. However, it might be necessary to have some exclusive information for products that would be useful for their proper disposal as waste.

The waste management architecture we have considered is built around several elements: waste items, domestic bin, trash bags, collective containers and collecting vehicles. The waste flow starts from the waste items and the domestic bin to end in the collecting vehicles. We now describe each of step in the waste flow and how these elements interact.

A. Wastes description

The presented management system is based on a self-describing approach of each waste. We have associated digital information to each waste to ensure an appropriate treatment of each item locally. This is the key point of this approach.

In the selective sorting process, the type of a waste item is identified by its main component. For example, a plastic bottle is identified as a plastic waste, and a cardboard box is identified as a cardboard waste. In the presented approach, each self-describing waste carried digital information about its type. Other properties of the waste are interesting for the collection process of the wastes. For example, the weight of each wastes can be used to estimate if a bin is full, or empty. Without using measurement sensors, the weight data of a waste item can be stored in digital information attached to it, making itself describing.

B. Wastes identification

The user is the primary actor in the selective sorting process. Based on this observation, our waste management system offers some pervasive assistance for the selective sorting process. Then, the waste flow presented in Figure 1, begins at the user level where the trash is generated. As it is shown on the top of Figure 1, we approach favors a behavior of the users: by indicating the appropriate bin for a piece of waste, or more directly, by opening the lid of the bin corresponding to the type of the waste.

C. Trash bag

To ensure an appropiate treatment, the knowledge of the type of wastes contained in a trash bag is crucial. As for the wastes, it is also possible to associate several properties of each trash bag: for example, the owner of the trash bag, and the number of items in the trash bag can also be considered. In the prototype presented in the next sections, some digital information about the total weight of the trash bag, its content and the number of items contained in the trash bag are physically associated to each trash bag. In this prototype, some digital information is also associated to identify the owner of a trash bag: the interest is to identify the waste production of each consumer. This information defines an analytical report associated to each trash bag.

The analysis report stores some important information for the selective sorting process. The information stored in the analysis report is to determine whether the trash bag could be accepted. In Figure 1, this analysis report is transmitted to the collective container, when a user brings a new trash bag.

D. Collective container

In our waste management system, each collective container is associated to an embedded computing system, which processes the data of the analysis report of each trash bag, making it a smart bin. When a new trash bag is added in a collective container, the analysis report is read.

Considering the type of wastes contained in a trash bag, a collective container determines whether it could accept a trash bag or not. For example, a collective container collecting only plastic wastes can stay closed when a user brings a trash bag containing the cardboard objects: it would only be opened for a bag of plastic wastes. If the trash bag is accepted, the smart bin stores some information about the content and owner of these trash bags. Then, the content of a collective container is iteratively updated as a new trash bag is added. The information stored by the collective container is transmitted to the truck during the collection by using a local connection, as it is presented on the bottom of Figure 1. At this step, the errors of the selective process can get transmitted. Among the collection of wastes, the highly polluting wastes,

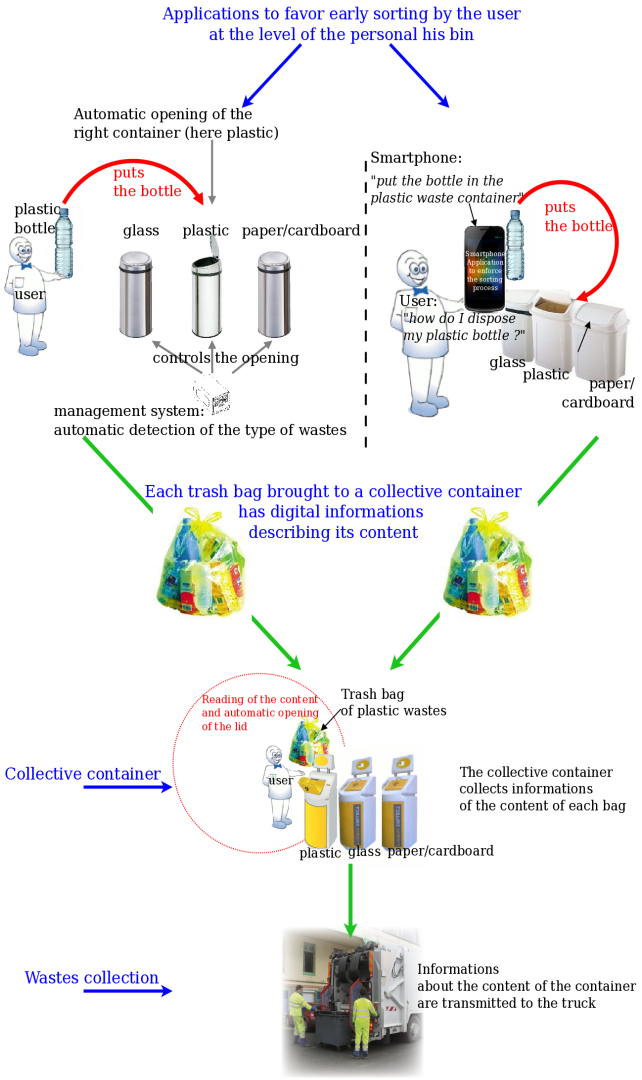


Figure 1. Waste flow and global architecture of the system.

which are not placed in the appropriate container, are detected: for example, it becomes possible to detect a battery placed in the container dedicated to plastic waste.

The focus in this step of the waste management chain is on the trash bags disposed by the users. On the one hand, it ensures on the individual waste items inside the bag without any tampering (contamination or removal); while on the other hand, it tries to channelize the bags in the proper recycling direction based on its dominant contents. This step can incorporate flexibilities; in the sense that compliance policies laid down by the waste management operator or the city could be enforced through these collective containers. A glass container could be made to accept trash bags containing glasses strictly without any contamination at all; or it could be also programmed to accept bags with leniency, i.e., having contamination upto a certain percentage and exceptions. In real life implementation there is need for tolerance as most of the waste is composed of various materials. Our solution to incorporate such flexibilities is described later in Section V. However, there would always be exceptional cases where a trash bag might not comply to any of the collective container. For such situations the waste management policy may provide a "catchall" bin that users can open with their personal identification card. In this way users could be tracked for either imparting recycling education (in case they are facing issues) or frequent defaulters not participating in the recycling program.

Considering this waste flow, we now present a system based on RFID technology to implement this waste sorting process.

III. Technological Support for Smart Waste Interaction

Our *smart waste approach* consists of associating a physical waste with digital information. In our approach, information associated to a waste item can be stored in a QR code or in a RFID tag memory. Using QR codes does not introduce an additional cost. However, QR code requires the object to be in line of sight. Unlike this technology, the RFID tags can be read without requiring a precise position relative to the reader during the reading operation. The UHF tags are used increasingly in the supply chain management and can be easily read at a distance of five meters from the reader antenna. In this context, it is easy to envisage a widespread deployment of the RFID tags on each manufactured product. This is an important advantage for using RFID technology in the waste management domain.

The tagged waste concept uses the data banks memory of a tag to store information about each waste associated to the tag. The tag memory is not used to store an identifier of the waste in an external database, but the information describing the associated waste is directly stored in the associated tag. Moreover, the tag(s) is(are) most likely to be placed on the "significant" part(s) of the waste to aid the user for better sorting. A connection to an external database is not required to have some information about the smart waste. Only a RFID reader is required to read the information of a smart waste. Figure 2 presents a smart waste composed of a plastic bottle associated to a RFID tag, which stores the data describing the bottle as a plastic object.

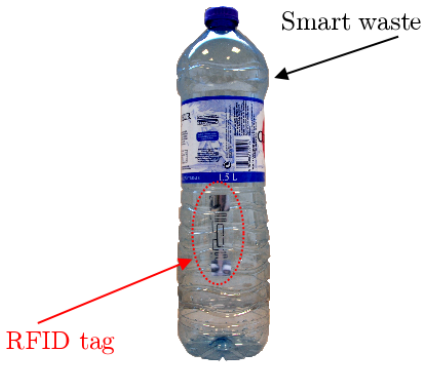


Figure 2. An example of a smart waste.

A RFID tag contains data banks for the users applications. The memory size of data banks is limited. For example, an UHF tag ALIEN ALN-9640 Squiggle shown in Figure 3 can store 512 bits of information.

Figure 3. The ALN-9640 Squiggle Alien tag.

In [6], the type of wastes classification is shown. In this classification, each type of waste is associated to an identification number. Taking examples from everyday life:

- the cardboard is associated to the reference 200101,
- the glass is associated to the reference 200102,
- the plastic is associated to the reference 200139.

The smart waste concept reuses the classification [6], to store the reference number representing the type of the waste in memory blocks of each tag associated to a piece of waste. As it is shown in Figure 4, our prototype also saves the weight (represented by a measure in grams, encoded in hexadecimal) of the waste associated to the tag, in the tag memory of each smart waste. The weight encoding presented in Figure 4 is a way to store the description of each waste directly in the associated tag.

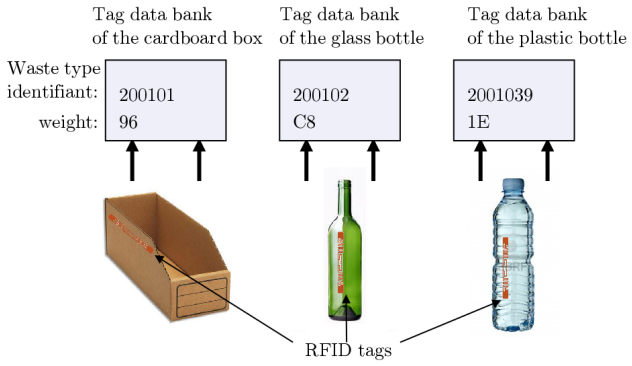


Figure 4. Representation of the information in the tag memory.

Until this point of the section we have seen how the digital information is associated to the physical waste objects. During their disposal at different stages of the waste management system, these information are read and transferred for processing and aggregation. The rest of this section describes the various modes of the domestic waste collection and the transfer of its digital information to the smart trash bags. Finally, the collective smart bin is demonstrated, which is suitable for community waste collection area. The smart trash bags are disposed here.

A. Individual smart bins

At the first step of our waste sorting system, the information contained in the RFID tag associated to each smart waste is used to help the people disposing an object in the appropriate container. Here, the main goal is to reduce the sorting errors when someone does not know, which is the right container, or mistakenly discards the object in the wrong one. It also helps people to learn the selective sorting rules applied locally. The smart bin system uses the self-describing approach of smart wastes to improve the selective sorting quality.

The description of smart wastes is stored in a RFID tag physically associated to each smart waste. Using a RFID reader, the smart bin reads the RFID tag attached to each smart waste to determine the appropriate treatment. Let us consider the example of someone who wants to discard a plastic bottle in a bin. He puts the bottle near a smart bin as it is shown in Figure 5. When the plastic bottle is in the antenna area, the tag associated to the bottle is detected. The data stored in the tag is read to determine the appropriate procedure to discard the bottle. If the bin accepts plastic objects, then the system opens its lid. Otherwise, the system keeps the lid closed.

Note that it is also possible to control the opening of several containers using a single RFID reader. Figure 5 presents a prototype of a selective bin. In this approach, a management system connected to a RFID reader uses the data stored on waste tags to open the correct containers. In this example, when someone wants to discard a plastic bottle, the container for the plastic wastes is opened by the management system. In Figure 5, only the lid of plastic container will be opened and all other lids will remain closed.

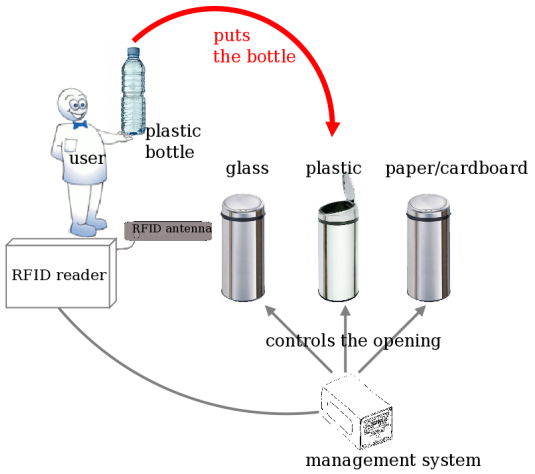


Figure 5. A RFID based selective bin.

This approach assumes that the management system tracks the information of the waste items that are discarded in each container of a selective bin. When a piece of waste is discarded in the container, the management system updates the memory inventory for this type of waste. In this way, undesirable wastes for a given container are either rejected or tracked, depending on the chosen policy for handling undesirable wastes. Products are scanned item-wise to ensure a complete reliable reading process. Figure 6 presents a prototype of a smart bin based on UHF RFID tags and a UHF RFID reader that implements this approach.

UHF RFID technology is already used in the supply chain

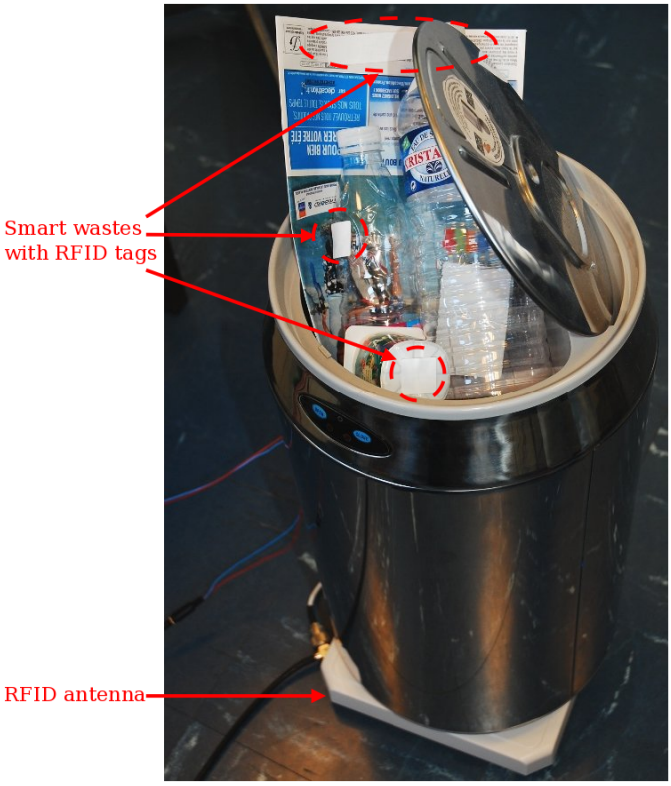


Figure 6. A collective bin using the RFID technology.

management systems. In this context, a UHF tag is placed on the packaging of each product at the begining of its life cycle. Since the UHF tag is already attached on the packaging of each product for the supply chain management process, we are interested to reuse the tag and its technology in our smart bin approach.

B. Use of QR codes technology for a cheaper approach

The passive UHF RFID tags are quite cheap, ranging from $0.10 to $0.15 per tag. However, the relative overhead cost and utility is important for businesses to adopt the technology. They would readily agree to tag an expensive item than the very cheap ones. To take care of this practical aspect, we introduce a cheaper solution using QR codes technology. It would allow an early adoption of few concepts and applications presented previously. Using QR code is not necessarily an alternative but could also be considered as a complement to RFID for cheap items. This approach takes advantage of the embedded NFC capability in users smartphone.

This alternative approach assumes that every waste is associated to a QR code describing its type. The mobile application maintains in its memory the current inventory for each type of collected wastes (for example, 3 inventories if there are 3 types of collected waste). Waste disposal would require users to scan each item, allowing the mobile application to update the current inventory for this type of waste in phone's memory. Some other waste properties, such as weight, could also be collected at this step.

A smartphone is a small, low-cost, mobile computer. Moreover, most smartphones now embed a camera, enabling them to read bar codes or 2-dimensional QR codes (also known as "flash codes"). A first step in the solution would consist to scan a QR code (or bar code) associated to a product, and to use this information for giving a sorting instruction to the smartphone of the user. As in the approach of the individual RFID bin presented in Section III-A, it is also important to report the actions of the user to the waste collecting chain.

In Figure 7, a user wants to drop a plastic bottle. He scans the QR code associated to the bottle. The properties associated to the bottle are added to the inventory of the plastic container that is stored in the smartphone's memory.

Figure 7. Reading of a QR code associated to a waste item.

Obviously, reading QR code is less convenient than RFID reading. Additionally, in this approach, the opening of the lid is also not controlled by an automated system. However, this approach allows the deployment of the rest of the chain without requiring the smart bins inside each home, as presented in Section III-A. Beside being cheaper, the mobile application also provides helpful support to the user regarding the selective sorting rules in application.

Like the individual bin presented in Section III-A, the management system of the collective bin tracks waste properties as they are disposed. When a smart trash bag associated to a RFID tag is dropped in the collective container, the management system updates the collective inventory according to the new bag's content. Prevention of sorting errors is also possible, provided that the user actually fills his trash bags according to what he scans.

We do not rely on a network connection of the bin. Instead, it is the waste bag itself that will store the waste inventory, as we will see in next section.

C. Smart trash bag

In the individual selective sorting point like a user's smart kitchen bin, the wastes are not directly deposited in the container of the bin. Every user utilizes trash bags, which will be dropped to a collective container in the residence, or put at the entrance of every household for being collected by the service provider.

The *smart trash bag concept* is smart in the sense that the waste management infrastructure (bin, truck) will be able to check its contents. A smart trash bag is a trash bag associated to a RFID tag, as it shown in Figure 8. The tag associated to a smart trash bag offers a memory space to store some information about the contents of the trash bag like: type of wastes, number of items, etc. The RFID tag may also store some information about its owner: name, address, etc.

Figure 8. A smart trash bag.

Writing data in the tag associated to the smart trash bag about its content is straightforward: for each new smart waste added, its tag is read; then, the trash bag content is updated by writing in its tag with the updated information about the newly added waste. This approach enables the tracking of trash bag content. Various information can be reported; i.e., the type and quantity of wastes contained in the bag, total weight of the content, and the interactions between the wastes. In this approach, it is assumed that the management system ensures that all the waste of a container belong to the same type. Then, it is just necessary to store the expected type of wastes in the analysis report. The weight of the smart bag is estimated by considering the weight of each smart waste contained in it. When a smart waste is added, its weight is read from the tag memory. The smart trash bag's weight is refreshed by adding the weight of this smart waste to its current weight. The weight is computed each time a smart waste is added. This iterative process uses the information stored in the tag associated to each smart waste. This approach is totally autonomous and based on the information stored in the tags associated to each smart waste. A connection to an external information system is not required to obtain the information associated to each waste.

As it is illustrated in Figure 9, the information stored in the tag associated to the smart trash bag, are encoded by a sequence of bytes. Storing the owner's identifier uses three bytes. Using the classification of wastes [6], the type of wastes is stored as six hexadecimal digits amounting to three bytes. The number of waste items is stored as one byte. The weight (in gram) of the content is stored as two bytes. Without requiring an external database, the description of trash bag's content is direcly carried by its associated tag.

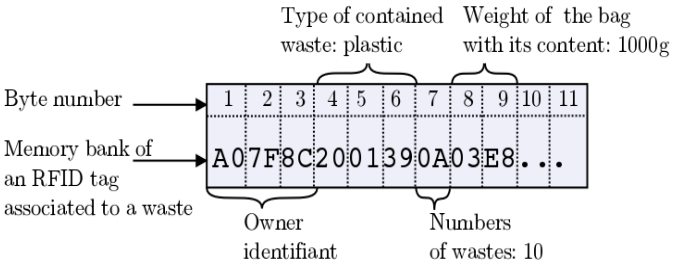


Figure 9. Data memory structuring of a smart trash bag.

The individual smart bin approach presented in Section III-A assumes that the management system tracks the information of the waste that are discarded in each container of a selective bin. To this end, the management system updates the analysis report of a container when a smart waste is added to the container. The whole report is stored locally in the memory of a tag associated to the smart trash bags. In the prototype presented in Figure 6, the analysis report stores the information about the owner of the smart trash bag, the type of the content, the number of wastes, and the weight of the content, using the data representation shown in Figure 9.

We have also developed an application to store the analysis report of the content of a trash bag for the cheaper solution using QR codes technology presented in Section III-B. In our prototype, the NFC technology provides this second step of the solution: NFC-enabled smartphones can interact in close proximity: in particular, they can read some RFID tags and also emulate the response of some tags. It is in the former functionality that we are interested, as it allows a user to write the required information in a trash bag's tag using only an NFC smartphone.

Figure 10. Writing analysis report operation with a smartphone.

When the bag for a given type is full, the mobile application is used to write the inventory in the RFID tag attached to the trash bag (Figure 10). The smartphone uses its NFC reader/writer for this operation. Then the user closes his trash bag of plastic wastes. Now, he uses his smartphones application to write the inventory of the trash bag, in an NFC tag associated to the bag.

D. Collective smart bins

The *collective smart bin* collects the smart trash bags produced by the users. Here, we consider a scenario for the collective smart bins, which can be placed in a common space of several apartments or in a street. Using the self-describing approach of the smart trash bag, the collective smart bin monitors the flow of wastes, and it detects the alerts like fire, sorting errors, detection of undesirable objects. The information about its content is transmitted by an ambient network or local Bluetooth connection during the collection, according to the type of information. As for the individual bin approach for helping the sorting process, it is possible to open a container only when objects of the correct type is brought by a user. The RFID inventories cannot ensure that all the tags have been detected in antenna area of a reader, meaning that missing tags are unnoticed. Considering this limitation, we have followed an "incremental" approach, where the global content of the collective container is updated each time a bag is disposed.

The analysis report of the content of a trash bag presented in Section III-A is used to update content of the collective bin. This approach is based on the self-describing concept of the content of a container; in the same way as the individual container stores knowledge about the wastes. The collective container stores knowledge about the smart trash bags. It is a new way to measure the state of a container without requiring the use of various sensors. For example, the weight of content, the size, or the type can be measured by using the information stored in the tag of each waste of a container, without using any sensors for each specific property. For example, the total weight of the wastes of a collective container can be estimated by incrementally adding the weight of each smart trash bag brought to the collective container. The information stored in the tag of the smart trash bag is only needed. This autonomous approach facilitates a large scale deployment of the smart bins.

Figure 11 shows a user in a garbage room. He presents his trash in reading area, where the trash bag's tag is read. The analysis report of the content of the trash bag is transmitted by the reader to the container's controller. The controller can then determine the appropriate action, depending on the bag's content and the local policy. For example, it could reject the bag if it contains an inappropriate item (container remains closed), explaining the cause of rejection to the user (such as "glass is not accepted in this container"). Implementing this policy is a way to avoid that a sorted container is contaminated by undesirable material. For example, it becomes impossible to pollute a container for plastic waste with metal cap of a plastic bottle.

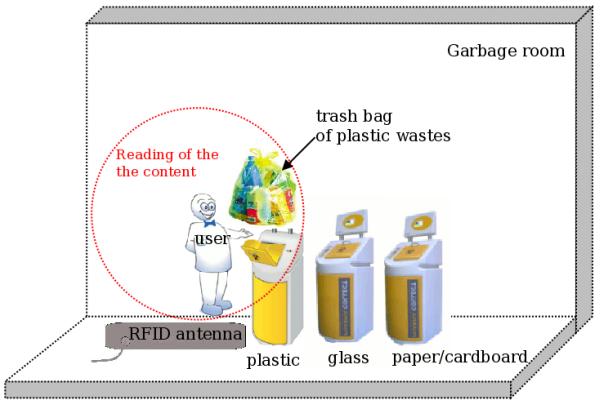


Figure 11. An individual bin using the RFID technology.

IV. An integrity certification process of the content of a trash bag for rewarding the selective sorting

Based on the smart waste and smart trash bag concepts, the applications presented in Section III, help users in the selective sorting process. Considering this selective sorting approach, it becomes easy to reward the behavior towards the environment. The smart bins use all the hardware required to implement a payback mechanism based on a micro-payment approach. Based ubiquitous computing principle, this can be implemented by a payback mechanism where the amount is credited into the user's smartphone. This non centralized approach makes deployment easier and offers better privacy for the user; therefore, not requiring a centralized server for storing data about the waste production details of each household. Irrespective of the implementation for payback mechanism, the reward process should be based on properties like the number of item(s), their weight, or the value associated to each waste. Thus, in our scenario, wastes and trash bags would have a value and they become critical objects for checking their integrity, mandatorily.

Our approach considers two steps of the selective sorting process. Firstly, the individual smart bin facilitates the selective sorting of the trash bag for the user. Secondly, the collective smart bin ensures selective sorting of smart trash bags, using their tag information. The collective smart bin uses the waste inventory stored in the tag of the trash bag, to ensure a real time waste management of its content.

In the waste management chain, some event might take place between the individual container and the collective container and disturb the selective sorting process. In particular, an undesirable object can be added; like for example, a battery can be added in a plastic waste trash bag. A mistake or a malicious behavior can corrupt the chain of the selective sorting between the individual smart bin and the collective smart bin.

Here, we suggest a method to implement a certification process of the content of a smart trash bag. Using a certification inventory mechanism presented in [7], the inventory of the set of all the waste contained in a trash bag can be

used and certified. This approach purposes to add a integrity information in a group of tags. Then, a RFID inventory of this certified group can be checked for consistency of the information distributed over the set of tags.

A. Certified content creation phase

In the selective sorting application presented in Section III, the certified content creation phase is made by a user with his personal individual bin. In the smart bin application, the integrity information is a hash value computed with the set tag identifiers at the level of the individual smart bin.

Considering a set of tags with unique identifiers $t_1, t_2, ..., t_n$. Each tag is associated to a piece of waste contained in a trash bag. The identifiers are ordered in a determined sequence (using a chosen order relation). Then, a hash function is applied to this information to compute the digest: $d = hash(t_1, t_2, ..., t_n)$. As shown in Figure 12, this hash value is used as a group identifier gid, stored in the tag of trash bag, which contains the set of wastes. This hash value is used as an integrity information, which enables the integrity checking phase.

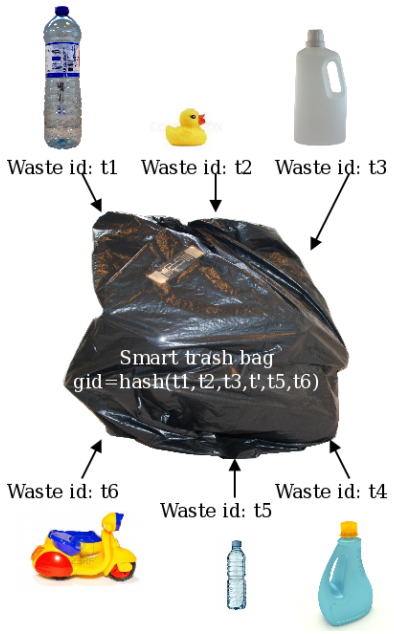


Figure 12. Building of a certified object set.

When a waste item is added in the smart individual bin, the management system stores the identifier in its memory. The management system of the individual bin incrementally stores the identifiers t_i of each waste added by the user. When the user closes his trash bag, the group identifier gid is computed with the identifiers $t_1, t_2, ..., t_n$ of each waste stored in the tag of the smart trash bag.

Because the memory size of the tag is limited and the integrity check should be fast, the group will be represented by a digest, computed by a hash code function. A good discussion of hash functions in the context of RFID is [8]. This approach enables full autonomous operation of both the association points and the checkpoints.

B. Checking integrity of content phase

The integrity checking phase is done at the level of the collective smart bin. The integrity checking phase considers the inventory of all the wastes contained in a trash bag. The integrity checking phase consists to verify if the hash value computed with all the tag identifiers read by the RFID reader, is equal to the integrity information stored in tag of the smart trash bag, at the individual bin level.

The principle is to read all the tags identifiers t_i of the wastes of a given trash bag (sharing the same group id gid), and verifying that the $hash(t_1, t_2, ..., t_n) = gid$. If the computed hash does not match the gid stored in the tags, the group of waste in a trash bag is considered as invalid. If wastes are removed from the trash bag, or if undesirables are put in it, the collective container will not open, because the integrity of the group of wastes inventory is violated.

RFID inventory would not require a line of sight with RFID tag for reading its contents: it is an important feature of the RFID inventory. This can be used to facilitate a checking process for the integrity of trash bag content, at the level of collective smart bin. It becomes possible to read all the tags associated to many smart wastes contained in a smart trash bag.

This behavior of the collective container may seem to be very restrictive. At the application level, some adjustments of behavior can be considered. For example, in the case where the integrity information of the smart trash bag is not valid, considering the set of all waste types, detected during the RFID inventory of the reader: if all the types of set are conformed to the type of wastes accepted by a container, the container can also be open. Using the RFID technology, this evolution of the system adds an automatic checking of the content of each trash bag.

C. Certification mechanism for the NFC and QR code approach

The certification mechanism can also be applied to the approach presented in Section III-B based NFC and QR code technologies. To do this, it is necessary to duplicate the information stored in the QR code of each waste to an RFID tag also associated to the same waste. The integrity information is computed using the identifiers stored in each QR code associated to a piece of waste, and it is written in the NFC tag associated to the smart trash bag using the smartphone application of the user. The informations contained in the tag will be used by the collective container presented in Section III-D. The collective container will check the integrity of the content of the trash bag, without rescanning the QR code of each waste.

The checking operation is to compare the group signature stored in the NFC tag of the trash bag, to the hash value compute with all the identifiers detected during the RFID

inventory. If they are equal, the content of the trash bag is valid, else the content of the trash bag is corrupted.

This certification mechanism using the NFC and QR code approach of the selective sorting remains cheaper, because the RFID reader are not deployed by the households, but only on the collective container.

V. Some more complex scenarios

Waste is an increasingly environmental issue for the society. If it is not disposed and treated properly, it can be detrimental to the living beings and the environment [2]. Managing the waste is a huge task, given its ever-increasing volume generated. They could even be complicated at times depending on the nature of waste. They come in many different forms like biodegradable, biomedical, chemical, clinical, commercial, electronic (e-waste), hazardous, industrial, nuclear, sharp, toxic etc. Each of the categories has to be processed differently. We refer to them as **complex objects** in this article. Hence, sorting must be performed at the earliest for performing appropriate treatment. However, as discussed earlier, some waste contains potentially useful materials for reuse, which are recycled. These need early separation through sorting, to prevent their contamination by other waste types and maximizing the amount of valuable recyclable materials contained in them. Hence, sorting is a very important process for waste management. In Sections V-A and V-B, we describe two other scenarios for efficient sorting.

A. Selective sorting

As described above, one of the aims for sorting is to maximize the amount of recyclable materials like paper, glass, plastic etc. We present another approach, which would enable this aspect of maximization. Smart waste containing the information about the amount of recyclables that could be recovered from them. Their tags are encrypted with the recyclable material classification type identifier and percentage information; instead of weight proposed, similarly, in Section III.

Consider the examples smart waste in Figure 13. The cardboard box, glass bottle and plastic bottle are made of recyclable materials. Each of their tags contain the information that they are made of cardboard, glass and plastic with 48%, 97% and 83% respectively. This kind of information would benefit in taking preferential decisions when sorting. Suppose a sorting process wants to gather glass (type identifier 200102) waste with atleast 85% purity. Among the three items in figure, the glass bottle in the center satisfies the conditions in respect of both, material and its quality. Hence, it would not be possible to contaminate a sorting process with inferior or other materials. As a matter of fact, contamination could also be the other way; like, adding the bottle from above example for a sorting process collecting low quality glass, between 40% and 75%. So, it is upto the sorting process to choose the purity range (in %) of recyclable materials while collecting waste.

Practically, there are various waste composed of multiple materials. A typical example of such form would be e-waste

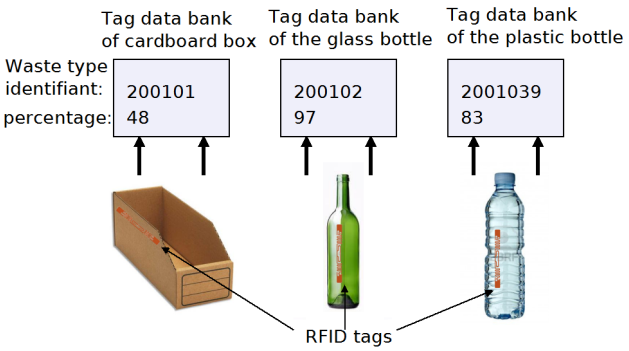


Figure 13. Representation of recyclable material percentage in the tag memory.

or electronic waste. Computers, telephones, televisions, etc.; all such electronic items contain recyclables and hazardous materials. Plastic, glass, metal are some that can be extracted for reuse. Hence, for such forms of waste, the above sorting process requires some modifications to its conditions. The sorting conditions must have flexibility to accept items containing multiple recyclable materials; unlike the process explained in the preceding paragraph. Consider for example a sorting process that accepts waste items containing glass (identifier 200102) $<$ 35% and plastic (identifier 2001039) $>$50%. The two conditions are represented by the two coloured circles in Figure 14 with their combined at the intersection. The monitor in the figure is tagged containing composition information of type and quantity of materials used, i.e., 25% and 55% of glass and plastic, respectively. The composition satisfies the conditions set for the sorting conditions and hence would be accepted.

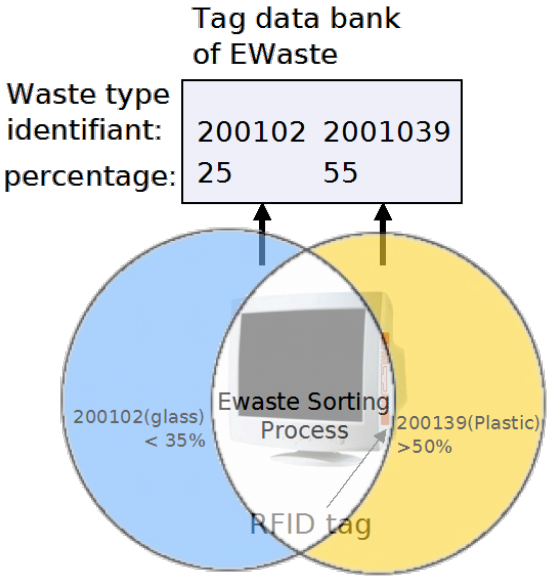


Figure 14. Representation of recyclable material percentage in the tag memory.

B. Detecting hazardous interactions

Until Section V-A, we explained how RFIDs attached to items store information regarding its quantative measures (weight and percentage) of recyclable materials enabling efficient sorting. The smart bin is a collective container that has an embedded computing system to read and process the tag data. However, there are other ways to perform sorting for better waste management. The sorting objective is to maximize on the concentration of recyclable materials for value. An accumulation could be contaminated due to the presence of other particular materials. This could render the entire collection unfit for recycling. Consider for example a glass bottle put into a paper or plastic bin. This would reduce the recycled value of the collected paper or plastic items [9]. Apart from mixing of materials, there could be physical hazards reducing the value of collected items. A flame caused by an explosion from aerosol can in high temperatures can ruin the collected paper or plastic materials. Disposal of such unsafe items in the same waste bin could result in a snowball effect of physical hazards. The remaining part of this section describes a sorting process that would enable avoiding such linked incompatibilities.

1) Principle: Self description of smart waste items contain information about their properties using RFID tags. Based on these properties, incompatibilities are computed among a collection of items present locally. In this section, we discuss its underlying principle. For the purpose, we begin with organizing the waste domain in a specific manner for making such inferences.

a) Describing waste items: The waste domain can be categorized based on their various hazardous properties. There are standards that specify the properties of waste materials and categorizes them [10]. Although, discussion on such standards is outside the scope of this paper, however, we utilize its idea for categorization and use few examples of hazards related to some of these categories.

Some examples of hazardous properties for this domain are spark, explosion, toxic fumes, etc. and can be categorized based on them. As discussed in the previous section, we are interested to infer incompatibilities. So, it is essential to pick the properties only that are relevant for interactions with other items.

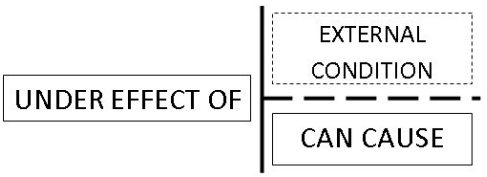


Figure 15. Conditions to describe a category.

Figure 15 represents pictorially the data structure used for describing waste categories. Its individual fields are described as follows:

- under effect of: the condition(s) that holds the properties that can influence the category
- can cause: this condition enlists the hazardous properties that the category is capable of causing
- in presence of: this holds the external conditions, under which the **can cause** properties occur; they are the physical environmental conditions that need to be captured using sensors.

In the subsequent sections, we will use the same pictorial representation to describe the waste categories or items in our examples.

Let us take some scenarios of interactions between categories. First, let us take an example of simple incompatibility between a pair of them. Suppose a category *A* can cause an incidence (for instance say hazardous property X) that affects a second category *B*. Hence, an incompatibility exists between the categories *A* and *B*. Our second example is a slightly more complex and realistic than the previous example. If the category *A* causes the incidence (i.e., X) only in presence of an external condition (let us name as C), makes it an important augmentation to the scenario. Hence, the categories does not pose to be incompatible if the condition C is unfavorable. Both of these scenarios consider the incompatibility between different categories where the hazardous property affects each other. However, there are properties like explosion for example, which have hazardous effect by itself. The situation can be represented as a category that causes a hazardous property that affects itself and may depend on the external condition.

b) Inferring incompatibilities: As described above, we can self describe waste items accordingly. When a collection of these items is present locally we can infer incompatibilities based on the discussed scenarios. Sometimes objects are located remotely and communicate within themselves and other knowledge base using network infrastructure like the Internet to make decisions. Such an idea is called Internet of things (IoT) in the field of pervasive computing. Our approach in this paper, makes the required information that describes waste domain available locally for inferences. Such collective inferences could be made without using a network for communication. We prefer to use the name for such a situation as Intranet of Things (InoT) as it does not involve any devices located remotely and differentiate to avoid confusion.

In Section V-B1a, we discussed the interaction scenarios between pairs of categories based on hazardous properties. Multiple such categories can constitute an InoT. The graph in Figure 16 represents an example of InoT formed. The shaded nodes represent some categories. They are connected by an edge if they interact. The dotted edges represent interactions that are unfavorable due to external conditions. One of the external conditions was high temperature at the instance this snapshot was drawn. Hence, the dotted edge encircled in the figure representing an interaction under low temperature becomes unfavorable. The firm edges represents favorable interactions, which could be either the first or second scenario described in Section V-B1a. The shaded node with a self-loop, which represents the last scenario of V-B1a, is favorable in this

case as the external condition is satisfied.

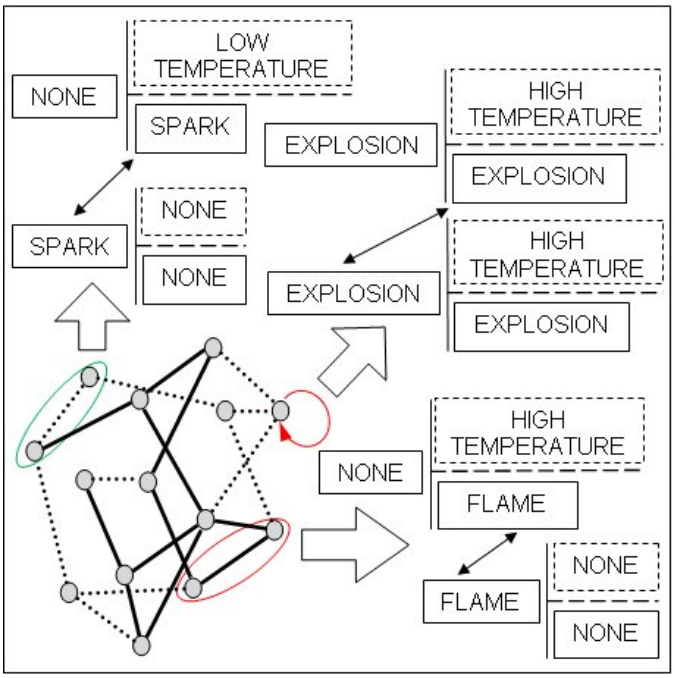


Figure 16. InoT formed.

Finally, if a waste item belongs to one or more categories, it would possess all their conditions. Hence, they could be used for collective inferences also.

2) System Design: In this section, we describe designing the system for making inferences locally. It essentially means that all the information required are available from self-describing waste without referring to remote database or knowledge base. An alternative could be to distribute the information partially among the waste items and a local knowledge base, containing the common domain knowledge. The waste items are identified by the system before inferring on incompatibilities. We have chosen a commonly used architecture for our system, as shown in Figure 17 below.

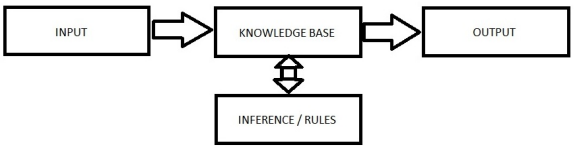


Figure 17. Commonly used Architecture for Systems.

We describe the components briefly.

- Input: It is that point in the system where the waste items are identified and added.
- Knowledge Base (KB): This contains all the required information to identify the items along with their properties. It also updates its knowledge regarding the presence of items that are being added to the system incrementally.
- Inference/Rules: This component of the model uses the KB to reason out about the possible incompatibilities and hazards. The inferences are added back to the KB.
- Output: It sends out notifications to communicate about alerts and warnings to the users of the system.

Next, we elaborate on how the system works based on the architecture and uses the principle discussed earlier in Section V-B1.

a) Input: New waste items are added to the system. They are affixed with RFID tags only for the purpose of identification by the system, which contains a RFID reader for scanning. The tags do not contain any such data that has privacy concerns. They contain mostly the category information.

b) Knowledge Base (KB): Machines can be made to perform reasoning effectively provided it has the necessary knowledge, which is machine readable. In cases of large domain knowledge with lots of factors influencing the reasoning, using machines should have extra benefits. Using ontologies are a very good way to serve the purpose [11]. An ontology consists of common set of vocabulary as shared information of a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them [12]. Lately, the development of ontologies has begun to find many uses outside the Artificial-Intelligence laboratories. They are being commonly used on the World-Wide Web and finds applications for sharing information widely in the field of medicine.

The Web Ontology Language (OWL) is a World Wide Web Consortium (W3C) Recommendation for representing ontologies on the Semantic Web [13]. Presently, there are a lot of ontology editors for OWL. Among them Protégé is a Java based Open Source ontology editor. We used Protégé since we found it to be an efficient and user-friendly tool to prototype our ontology rapidly. During the ontology development phase we visualized the graphical representation of our OWL ontology on the editor. The comprehensive Java API provided by Protégé [14] was also an added advantage while developing our stand-alone application in the later phase.

We have used an ontology based approach for the KB for the reasons stated above. The properties causing incompatibilities must be described in the ontology. Apart from these, other information like conditions in which the categories are incompatible, possible hazards of incompatibility etc are also stored in the ontology.

Due to the advantage for describing a domain easily, we have used ontology based approach for describing the waste domain. The ontology contains description of various categories with the conditions for hazardous properties. This constitutes as the initial knowledge base of the system, which is maintained locally. It updates itself as new items are added. Additionally, the external conditions are also updated from the environmental parameters from sensor data. The modelling and design of the ontology is detailed in the paper [15].

c) Reasoning/Rules: Reasoners are a key component of OWL ontologies. They are used for deducing implicit knowledge by querying the ontology. In the recent years, rule

languages have been added on as a layer combined with ontology, in order to enhance the reasoning capabilities. Semantic web Rule Language (SWRL) is used to write rules expressed in terms of OWL concepts and for reasoning about OWL individuals. It provides a deductive reasoning specification that can be used for inferring new knowledge from the Knowledge base.

The ontology, which acts as a KB in our architecture, contain all the necessary information for reasoning. The principles described in Section V-B1 for detecting incompatibilities between categories of waste items are implemented as ontology rules. Our objective of inferring incompatibility or hazards based on these rules are performed using OWL reasoners. The reasoner springs into action each time the RFID reader detects a tagged item. It infers if the incoming item has incompatibility with the already present contents using the ontology KB. The reasoner also provides the analysis, if found unsuitable.

3) Applications: In this subsection, we describe the system using ontology as its local knowledge base to infer incompatibilities on the principle of InoT. We think that it can be used to infer incompatibilities among objects in various domains. "Bin That Thinks" is a project, that is designed to have an intelligent waste management solution based on item level identification. The goals are to improve recycling efficiency, reducing waste processing cost and avoiding hazardous situations [16]. Though we have not assessed the financial benefits figuratively for using our system, the approach hints at the benefits qualitatively. Sorting waste items at the earliest retains the purity of the recyclables. This reduces the cost of sorting at a later stage in processing plants by waste management companies like Veolia, which is usually passed on to the consumers as penalties of the cities.

Figure 18. Final Smart bin Prototype.

We have developed an application for the domain of waste management using the system described in this paper. It can be used to make inferences for incompatibilities and hazards among the waste items present collectively at a place. They may be situated inside a bin or a waste collecting vehicle or at the processing plant. For very complex domains like waste management, they are sometimes verified at every step in the processing chain. Alternatively, when the processing is performed at a single point, we consider the acceptance of error up to some limit. Figure 18 shows a prototype of the final Smart bin that would identify the RFID tagged wastes and make inferences from its contents. It contains an RFID reader, an on-board processor, environment sensors for temperature, pressure etc., an OWL ontology based knowledge base and a display. The reader senses and reads the category when an item is brought near the bin. Then the reasoner makes required inferences using the KB before the bin actuates. The appropriate lid opens for the user to dispose the item, if it is found suitable with a green signal on the display. Else, the display flashes red along with the reasoning for incompatibility. Figure 19 below shows a screenshot of our application. It shows the instance when an incompatibility is detected with two items present locally in the bin and the last item that was scanned. It also displays the inferred reasoning.

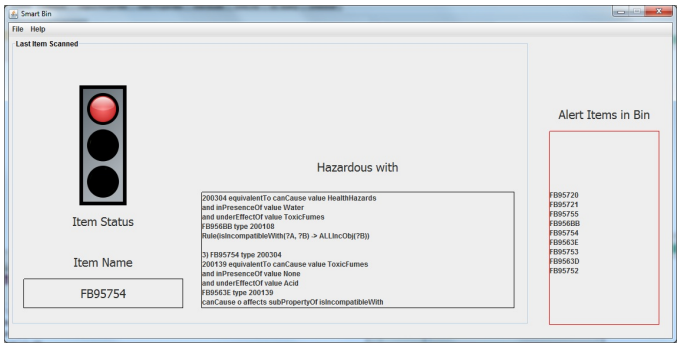


Figure 19. Hazard Detection Application for Waste.

VI. Communication Architecture Overview

The self-describing information collected at the level of the collective containers are useful for the recycling service provider. These information can used to enhance two aspects of the recycling process:

- to optimize the waste collection scheduling by the truck (it is not necessary to collect an empty container) and monitor their waste inventory
- to adapt the treatment of the content of the bins (if a container is polluted by an undesirable product, which requires to have a specific treatment). For example, when a bin collecting recyclable plastic has been polluted by glass; requires careful handling during collection.

Obviously, it is necessary to communicate this information before the waste collection. The self-describing information of each collective container has to be transmitted to the recycling service provider's terminal, or to the truck driver collecting waste from the bins.

The communication process between smart collective containers is based on multi-hop wireless networks, as it is shown on the top of Figure 20. In this communication architecture, each collective container is a starting point of the multi-hop network. The data of a given smart collective container are routed from the smart container to the recycling service provider's terminal (or to the truck that collects waste from the bins). So, the data hops through several communication nodes before being received by the terminal. This architecture (similar to Wireless Sensor Networks) is possible due to the urban topology where each bin is very close to the others. This non-centralized approach is less costly (financially) than using a GPRS connection between each collective container and the service provider's terminal. Also, each container's energy lifetime is very crucial and a balance among them spread across the city is maintained [17].

The motive behind incorporating this feature enables efficient waste management. The different types of waste collected by the operators have associated monetary value. Hence, they require sorting based on their types before being sold to third parties who recycle and reuse them. While the sorting is taken care at the bin level, the communication infrastructure of bins help in efficient collection and marketing. The operator can have a global view of the current stocks of the by-product materials (plastic, glass, etc.) in the bins for a given city, or even country-wide. Finally, it can receive alerts in case of incompatibilities or physical hazards in the waste containers, as described in the previous section.

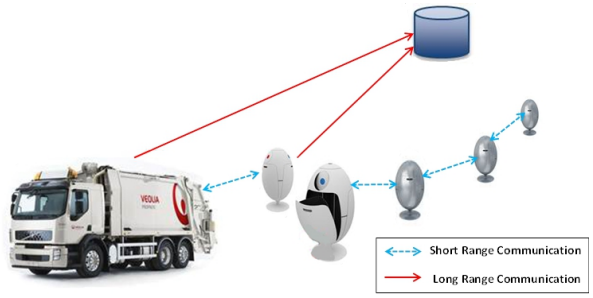


Figure 20. Communication architecture.

Partners of the "BinThatThinks" project [16] have implemented this ambient network with nodes offering an effective radio-frequency communication. In a city, it is difficult to deploy a totally reliable ambient network of bins, due to environment constraints. In fact, the communication graph (where each vertex is a communication node, and each edge is a communication link between two nodes) can not be connected. Then, this architecture requires some communicating nodes to have a GPRS connection for sending data from parts of the ambient communication network, which are not directly accessible to go through the ambient network. These nodes are more costly, and use more energy. An energy efficient protocol for long life operation such as waste containers is presented in [17]. This protocol maximizes the combined battery life of the global infrastructure. Using this protocol, all the batteries have to be replaced at the same time. To achieve that, it uses an energy balancing system. This aspect is important for the support of the network. To reduce the maintenance cost, it is particularly interesting to fix a replacement date of the set of batteries of all nodes of the ambient network.

VII. Prototype Demonstration

Figure 21 shows the prototype demonstration developed out of the collaborative project "Bin That Thinks" [16]. It consists of prototypes developed for the various solutions of the waste management chain discussed in this article and are marked numerically. The first shows two kitchen bins that collecting different types of waste. The QR codes are scanned using the smartphone placed in-between, which is also used to write the inventory onto a NFC tag when a trash bag is sealed. The second is the collective container that can scan the inventory tag and verify with the bag contents for compliance. Its optional screen displays the status and statistics. The later two shows the required applications deployed for monitoring in the truck and operator's station, respectively. While the former visualizes information about bins to be collected, the later displays the waste management infrastructure over the city map. Appropriate operations could also be done through these applications. Finally, the entire demonstration can be viewed in [18].

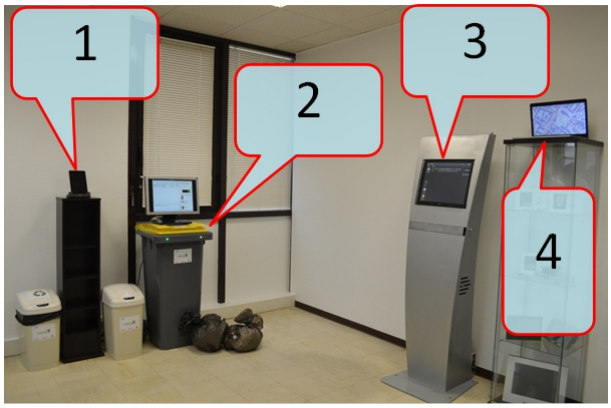


Figure 21. Prototype demonstration: (1) Kitchen bin (2) Collective bin showing status (3) Truck's Application Screen (4) Operator's status application at city level.

VIII. Related Work

Some other approaches using the RFID technology for waste management systems does already exist. In [19], the author discusses about several applications of the RFID in the process of collecting wastes. The identification of each bin associated to a RFID tag is mainly mentioned. The tag memory associated to each product has information about its end-of-life management process; it is also illustrated in the paper.

These approaches describe methods to identify each bin using an identifier stored in a RFID tag associated to the

container. In [20], [21], using this identifier, the author associates each container to an external database, along with the address of the household owning the container. The volume of wastes estimate the quantity of waste produced by each household. It is not an information based approach but a physical measurement approach using sensor. The estimation of the volume of waste is computed using an image analysis from a picture of the content of the bin (when the lid of the container is opened). The data are transferred using a GPRS connection to an external database.

The idea developed in [22] is also very close to this approach. In [22], a sensor measures the weight of the bin placed on the truck, which collects the bins. It differs to our approach, which uses a self-describing approach of wastes to compute the weight of a container. At each collection operation, the truck saves the weight of each bin. The RFID tags are used to store an identifier in a external database of the owner of each container. This approach is not autonomous, but using a Wifi connection, the external database of waste production is updated for each household. It becomes easy to track the waste production of each household. In [23], the author presents a real deployment of a system using an approach similar to the approach described in [22].

The concept developed in [24] rewards consumers for recycling empty packaging. The consumers are identified by a RFID tag associated to their bin. Based on a weight measurement of their recycling packaging, the consumer can also log into his personal account to view how much they have recycled, as well as statistics such as the number of trees saved by their effort. Every month, the consumers are also rewarded financially.

Actually, selective sorting is not the priority of these applications. RFID is used by the container to identify its owner. To ensure the selective sorting, it is required to track waste at the item level. This is why item level RFID tagging can have an important role to play in the selective sorting, provided that the tag contain information about the components of the waste.

In [25], the presented approach also considers that each product is associated to an RFID tags from the begining of its life cycle. The information stored in the RFID tags is not used to help the user in the selective sorting process. The authors use the RFID technology at another level of the selective sorting process. The RFID tags associated to each product is used to help the recycling service provider to decide about the appropriate treatment of the product. In this approach, the data stored in the RFID tags are used to access the products' information in several databases from its single identifier. This approach of using the RFID technology in the recycling process is not autonomous. A major difficulty is then to share conformable information about a product across several databases, and during all the life cycle of the product.

We had presented a complex scenario where interactions between various waste categories are inferred. A similar ontology based model is presented in [26]. Its application domain is in delimited environments where objects are located and is used to statically prevent/detect their dangerous spatial/temporal configurations. Although it could be an alternative for our case, the preciseness and complexity is on the higher side to be used when considering our real-time scenarios; like a user waiting with a smart trash bag for the bin to infer and open the lid for disposal.

The main goal developed in [27] is to bring out the environmental impact of RFID used in everyday life. The author discusses RFID for the waste management: a system of discounts and fees to stimulate responsible behavior of users in the selective sorting process, is also discussed. The idea of a bin, which collects some information about the wastes is mentioned, although its implementation is not discussed.

Although RFID tags has started to find widespread usage, it is yet restricted to certain applications due to limitations from the technology aspect. The reading reliability of tags vary due to certain conditions; like for example a lot of tags placed close to each other for reading. There are ongoing research at many places. [28] is one of them that aims to identify the challenges and propose solutions for better RFID usage in pervasive computing.

More generally, in [29], the author predicts an important development of RFID applications in the product recycling chains.

The approach that we presented in this article, is innovative in its information processing architecture: the properties are directly attached to physical objects (waste, bags) and data are "moved" and processed along with the physical flow of wastes. Several systems for encoding the waste description are discussed. The most simple way is to encode the component of each waste in plain text.

Value addition

We presented an architecture that is novel compared to the existing literature in the best of our knowledge. We have demonstrated it through various use cases (smart waste and trash bag, collective bin) in the context of waste management domain. [30], [31] elaborates the same with other use cases. Ours has the capability to perform operations autonomously, unlike the current approaches that requires centralization for either information or its processing. Its benefits are in terms of the following:

- cost cutting - Deployment and usage of industrial network, required for such purposes is expensive. The cost also reduces with having minimal number of centralized servers. Also, we have proposed the reuse of the existing RFIDs attached onto objects by manufacturers, retailers etc. Self-describing the waste does away with the installation of sensors [22], [32].
- scalability and availability - Our approach has high scalability and availability due to local processing and self-description. Consider for example a scenario, where messages are transmitted over the network for information and processing everytime a waste item is disposed; thus dropping the scalability drastically. The working of such a system would breakdown due to the unavailability of

the servers and/or network.

- privacy - The information is aggregated naturally in our architecture. This limits the users' minute personal information to reach the operator's centralized servers, in the waste flow chain. It would be more acceptable for a user that the operator knowing the total glass/plastic waste he produces than the number of coke/juice bottles disposed.

Hence, from the research perspective, our architecture clearly demonstrates benefits as well as novelty.

IX. Conclusion

In this paper, we demonstrated a new solution to enhance waste collection efficiency using the RFID technology. Fully relying on digital information attached to waste items, this approach does not require any sensor, nor external information system support, enabling high scalability, availability and privacy. The presented system helps the user in correctly sorting and disposing wastes.

Regarding the user-support provided during waste disposal, he is directed towards the proper container for better sorting, and is helped in case of errors. We presented two approaches in this article; first for simple waste composed of one principal material, and the second for more complex waste composed of several materials. Another contribution of this system is to be able to report the contents of a bin. This information is useful for waste processing operators, for example to optimize waste collection scheduling, or to set up a special handling when an undesirable product is detected somewhere. This information is communicated to the operators using an ambient communication network of smart bins.

The reported information about the content of each bin is also a way to compute statistics of each type of waste in the recycling process. The smart bins can precisely determine the quantity of each type of waste produced by a household. It should help people to contribute to a more efficient sorting of waste, and reuse valuable materials. By considering the value of wastes produced by each household, it becomes possible to make a retributive incentive system to encourage each user to make the selective sorting of its wastes. This approach can also help to plan waste collection in better ways and with provision for operator interventions, in case of abnormal conditions.

References

[1] Y. Glouche and P. Couderc, "A smart waste management with self-describing objects," in *SMART 2013, The Second International Conference on Smart Systems, Devices and Technologies*, 2013, pp. 63–70.

[2] "The EUs approach to waste management," April 2012 (accessed 17 May 2015, 21h13). [Online]. Available: http://ec.europa.eu/environment/waste/index.htm

[3] H. Boileau and H. Björk, "Comparing household waste treatment policies between two medium size cities: Borås (sweden) and chambéry (france)," in *Proceedings of the 7th World Congress on Recovery, Recycling and Re-integration*, June 2006. [Online]. Available: http://csp.eworlding.com/3r/congress/manu_pdf/420.pdf

[4] "Veolia: Research and Development," (accessed 17 May 2015, 21h129). [Online]. Available: http://environmentalpassion.com/Research_and_development

[5] "Better Sorting for Better Recycling," (accessed 17 May 2015, 21h39). [Online]. Available: http://environmentalpassion.com/resource.php?id=2239

[6] Assemblée des Chambres Françaises de Commerce et d'Industrie, "Classification des déchets," November 2011 (accessed 17 May 2015, 21h10). [Online]. Available: http://www.enviroveille.com/public/documents/nomenclaturedechets.pdf

[7] Y. Glouche and P. Couderc, "A robust RFID inventory," in *Proceedings of the European Conference on Smart Objects, Systems and Technologies (Smart SysTech 2012)*, Munich, Germany, june 2012.

[8] M. Feldhofer and C. Rechberger, "A case against currently used hash functions in rfid protocols," in *Proceeding of RFID Security, RFIDSec'06*, 2006, pp. 372–381.

[9] "A way forward for glass recycling," (accessed 17 May 2015, 21h30). [Online]. Available: http://www.waste-management-world.com/articles/print/volume-10/issue-1/features/a-way-forward-for-glass-recycling.html

[10] "European waste catalogue and hazardous waste list," (accessed 17 May 2015, 21h31). [Online]. Available: www.environ.ie/en/Publications/Environment/Waste/WEEE/

[11] M. Cambillau, E. El-Shanta, S. Purushotham, and G. Simeu, "Owl ontology for solar uv exposure and human health," *Advances in Semantic Computing, Eds. Joshi, Boley & Akerkar*, vol. 2, pp. 32–51, 2010.

[12] T. Gruber *et al.*, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[13] "World wide web consortium (w3c), "owl web ontology language overview"," (accessed 17 May 2015, 21h32). [Online]. Available: http://www.w3.org/TR/owl-features/

[14] H. Knublauch, M. Horridge, M. Musen, A. Rector, R. Stevens, N. Drummond, P. W. Lord, N. Noy, J. Seidenberg, and H. Wang, "The protege owl experience," in *OWLED*, 2005.

[15] A. Sinha and P. Couderc, "Smart bin for incompatible waste items," in *ICAS 2013, The Ninth International Conference on Autonomic and Autonomous Systems*, 2013, pp. 40–45.

[16] "Bin That Thinks," (accessed 17 May 2015, 21h33). [Online]. Available: http://binthatthink.inria.fr

[17] D. Tony, N. Mitton, and M. Hauspie, "Energy-based Clustering for Wireless Sensor Network Lifetime Optimization," in *WCNC 2013, The Wireless Communications and Networking Conference*, 2013.

[18] "Bin That Thinks prototype demonstration video," (accessed 17 May 2015, 21h34). [Online]. Available: http://www.youtube.com/watch?v=daac0tFig34

[19] S. Abdoli, "Rfid application in municipal solid waste management system," in *IJER International Journal of Envionment Research*, vol. 3, no. 3, July 2009, pp. 447–454.

[20] M. Arebey, M. Hannan, H. Basri, R. Begum, and H. Abdullah, "Integrated technologies for solid waste bin monitoring system," in *Environmental Monitoring and Assessment*, vol. 177. Springer Netherlands, 2011, pp. 399–408.

[21] M. Hannan, M. Arebey, H. Basri, and R. Begum, "Rfid application in municipal solid waste management system," *Australian Journal of Basic and Applied Sciences*, vol. 4, no. 10, pp. 5314–5319, October 2010.

[22] B. Chowdhury and M. Chowdhury, "Rfid-based real-time smart waste management system," in *Proceedings of the Australasian Telecommunication Networks and Applications (ATNAC 2007)*, December 2007, pp. 175–180.

[23] J. Wyatt, "Maximizing waste efficiency through the use of rfid," April 2008, Texas Instruments Incorporated. [Online]. Available: http://www.ti.com/rfid/docs/manuals/whtPapers/wp_lf_hdx.pdf

[24] C. Swedberg, "Rfid Helps Reward Consumers for Recycling," in *RFID Journal*, February 2008. [Online]. Available: http://www.rfidjournal.com/article/view/3936

[25] A. Parlikad and D. McFarlane, "Rfid-based product information in end-of-life decision making," in *Control Engineering Practice*, vol. 15, no. 11, 2007, pp. 1348–1363. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967066106001535

[26] D. Cacciagrano, F. Corradini, and R. Culmone, "Resourcehome: An rfid-based architecture and a flexible model for ambient intelligence," in *Systems (ICONS), 2010 Fifth International Conference on*, April 2010, pp. 6–11.

[27] V. Thomas, "Environmental implications of rfid," in *Proceedings of the 2008 IEEE International Symposium on Electronics and the Environment*, ser. ISEE '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 1–5.

[28] "Pervasive rfid," (accessed 17 May 2015, 21h40). [Online]. Available: http://www.pervasive.cominlabs.ueb.eu

[29] D. C. Wyld, "Taking out the trash (and the recyclables): Rfid and the handling of municipal solid waste," in *International Journal Of software Engineering & Applications (IJSEA)*, vol. 1, no. 1, January 2010, pp. 1–13.

[30] A. Sinha, "Self-describing objects with tangible data structures," Ph.D. dissertation, Université Rennes 1, 2014.

[31] "Self-describing objects with tangible data structures," (accessed 17 May 2015, 21h42). [Online]. Available: http://videos.rennes.inria.fr/soutenanceArnabSinha/indexSoutenanceArnabSinha.html

[32] A. F. Silva, G. L. Sandri, L. Weigang, R. L. de Queiroz, and M. C. de Farias, "Irsom: Automated sorting of the waste from selective collection using content based image retrieval and self-organizing maps."

Making Context Specific Card Sets - A Visual Methodology Approach

Capturing User Experiences with Urban Public Transportation

Alma Leora Culén and Maja van der Velden
Department of Informatics
Design of Information Systems
University of Oslo
Oslo, Norway
{almira, majava}@ifi.uio.no

Abstract - **The paper discusses the use of visual methodologies in the sense-making phases of Human Computer Interaction (HCI) design processes. The discussion is illustrated through development of a card set, a visual tool, to explore context specific issues related to experiences with urban public transportation. The card set was intended for an open exploration of users' experiences during different phases of a typical commute, from preparing for traveling to arriving at the destination. The paper argues in favor of increased use of visual methodologies in HCI and presents a framework for visual methodology in the production of a card set. The framework consists of seven concepts that support visual reasoning: visual immediacy, impetus, impedance, association, abduction, blending, and analogy. Our results show that these concepts were useful for finding out what types of images were communicating precisely the intended meaning and what types inspired associations, blending, and abduction.**

Keywords - visual methodologies; visual methods; card sorting; service design; experience ecologies.

I. INTRODUCTION

This paper extends our previous work on making and using cards for capturing user experiences in public transportation [1], in which we described how a card set intended to capture user experiences with urban public transportation was designed.

Card sorting is a simple and frequently used method in human-computer-interaction (HCI). There are many examples of how and when cards are used in order to provide structure and guidance to design processes, e.g., [2], [3]. While card sorting is one of the basic tools in HCI, the making of card sets for an open, or semi-open, exploration within specific contexts is less frequently discussed.

This latter point, in conjunction with increasing popularity of visual methodologies as an epistemological tool in anthropology and social science [4], inspired us to look deeper into the use of visual methodologies for making card sets.

This paper, then, extends the previous work by focusing on concerns of methodology: 1) how to design a context-specific, card set usable in start phases of participatory design processes, i.e., for group-based sense-making based on open or semi-open sorting; 2) how to make use of visual methodology in the process of the card making; and 3) what are useful concepts that support the use of visual methodologies for card sorting.

The paper is structured as follows. In Section II, we provide some background on card-based tools and their use in HCI. In Section III, we discuss visual methodologies and visual reasoning, introducing the framework consisting of seven concepts that support visual reasoning. In Section IV, we discuss the service experience design, which is the context for our specific case of card set development. Section V shows how we developed the cards for use in participatory, context-driven workshops exploring user experiences in public transportation. In Section VI, our findings are presented and discussed. Section VII concludes the paper and addresses future work.

II. THE USE OF CARDS IN HCI

Card sorting is a knowledge elicitation method, initially used to find appropriate categories in the design for the web [5]. The cards typically represented menu entries and hyperlinks, and users were asked to sort them into meaningful categories. The cards could be produced manually, or automatically using software such as that in [6].

Wölfel and Merritt provide a survey of often used card-based design tools [7]. They analyzed their use and found five categories that highlight differences among the attributes of the various tools. The categories included the intended purpose and scope for the tool, duration of use and placement in the design process, methodology of use, customization, and formal qualities. Furthermore, the tools were classified as generic, customizable, or context specific. The purpose of the cards may vary from explorative, inspirational purposes challenging designers to think in another way, to an inquiry into a very specific use context. They may be utilized in any phase of the design process, for quick insights or deeper context inquiries through, for example, workshops. The use methodology has to do with how the cards are used and results analyzed. 'Open sort' does not impose any rules while the 'closed sort' uses pre-determined categories. There is also an in-between variant, semi-open sorting, where some suggestions on how the cards are to be used are given. Furthermore, the cards may

not allow any customization, allow optional, or require full customization in order to be useful. The latter is often the case when working in a specific context, such as the service experience design for a particular type of service.

The last category, formal qualities of cards, enables researchers or designers to determine the aesthetics (colors, moods etc.), visual appearance of cards (type of representation featured on the card, such as images, graphs, text), and physical qualities (e.g., size, material, and texture). In addition, formal qualities also encompass concerns related to how many cards one should have, if there should be multiple cards, or multiple sets of cards, different categories of cards, and so on.

We have used many card sets in own work, in diverse design settings and purposes, such as future workshops, explorative workshops [8], interviews [9], etc. IDEO [10], PLEX [11] and Design with Intent [12] cards were among favorites, in particular, for reminding users of diverse design and evaluation methods. For this research, though, AT-ONE service design cards [13] were particularly inspiring.

In addition to card-based tools, there are good resources explaining how to use the cards. For instance, Spencer [14] shows how to plan and run a card sort, analyze the results, and apply the outcomes to various projects. The book has also a chapter dedicated to making of cards. However, the approach presented is different from the one we present in this paper.

The purpose of the cards made through this research is to inspire, re-imagine and inquire into experiences within the context of urban public transportation. The cards needed to be made in a way that best facilitates the inquiry. But it is the method of making them that is the main outcome of the research. Thus, visual reasoning and visual methodologies were seen as helpful to the endeavor.

In the next section, we provide a short background on visual methodologies and what makes them now into accepted research methodology.

III. VISUAL METHODOLOGY

Visual methodologies are becoming more acceptable, and central, in research within social sciences and humanities. Several books on visual methodologies, such as those by Rose and Pink [15]–[17], were recently published. They advocate the need for better understanding and further development of visual methodologies and consider the research within the field as "*an area of academic and applied research that demonstrates particularly powerfully that the relationship between theory, technology and method should not be separated"* [16, p. 3].

The link between the technology and visual methods has become highly relevant with the widespread use of mobile phone cameras that enable easy production of large amounts of visual material on one hand, and general availability of technological platforms that support search, manipulation, design, and analysis of visual contents on the other hand. Researchers, thus, have powerful tools at their fingertips that enable them to make sophisticated decisions based on visual materials. These, in turn, support the emergence of new theories on how the new knowledge emerges when using visual tools and materials.

A. Visual methodologies, methods, and HCI

Methodology is concerned with comprehending how research is done and how chosen ways of doing it (methods, tools and techniques) lead to knowledge production. Methodology is also concerned with how the environment in which it is applied supports knowledge generation processes. According to Rose [17], visual material is always embedded in the social world and can only be understood when that embedding is taken into account. In HCI, technology becomes part of this relation, leading to both social and technological embedding.

In HCI, images, video, sketching, drawing, paper prototyping, and card sorting have long been used as visual methods, tools or techniques, e.g., participatory video [18], card sorting [2], Photovoice [19], collaborative drawing [20], photo-documenting and visual ethnography [21], [22]. However, a Google Scholar search with keywords "visual methodology" (ies) AND HCI does not give many relevant results, indicating that, perhaps, there is room within HCI to discuss visual methodologies, both from a theoretical and a practice perspective.

Thus, while visual methods are widely used, visual methodologies are not widely discussed. The two terms, method and methodology, are often, inaccurately, taken to mean the same thing. For example, the card sorting is a visual method for achieving some pre-set goal. Visual methodology has to do with principles that guide research practices, how the research with card sorting is, or could be done. It is concerned with questions such as how are images that are used on cards made, understood and interpreted, how are cards put in use, and how one generates knowledge through their use.

In line with Pink [16], we consider visual materials to be part of the knowledge-producing processes in HCI, in which methods that inform the process, tools, people that created the visual material, technology that supports its use, and research aims cannot be separated. This stance is also in accord with theoretical shifts within HCI, towards more qualitative research including phenomenology, senses, ecologies, experiences, and practices [23], [24].

B. Visual reasoning

It is often said that images (and other types of visual-spatial materials) augment cognition [25]. In [26], Hegarty provides arguments from cognitive science as to why this is so: 1) images are external representations, freeing the working memory for other aspects of thinking; 2) grouping related information is a natural property of perceptual organization, often reflecting Gestalt principles; 3) allows for the offloading of cognitive processes onto perceptual processes; and 4) with interactivity, people can offload internal mental computations on external manipulations of images.

In proposing a visual methodology for making card sets for inquiry into specific contexts, we propose a framework consisting of seven concepts that support visual reasoning.

These seven concepts are immediacy, impetus, impedance, association, blending, analogy and abduction. They all support visual reasoning and augment cognition utilizing one or a combination of arguments presented above. We now provide definitions and use examples.

In [27], we discussed the role of information visibility in public transportation ticket systems. By contrasting the visibility of the ticket information of paper tickets and smart card tickets, we found that what people really appreciated about paper tickets was the availability of ticket information 'at a glance', i.e., the paper ticket had *visual immediacy*. What they liked the least regarding smart cards was the lack of such information. Much of our design efforts consequently focused on how to design for visual immediacy in the smart card based transport system. Diverse issues related to visibility of ticket information were considered, and diverse solutions proposed. An augmented reality application for a smart phone was prototyped, with which one could see all the information stored on the smart card 'at a glance', just like on the paper ticket, see Fig. 1.

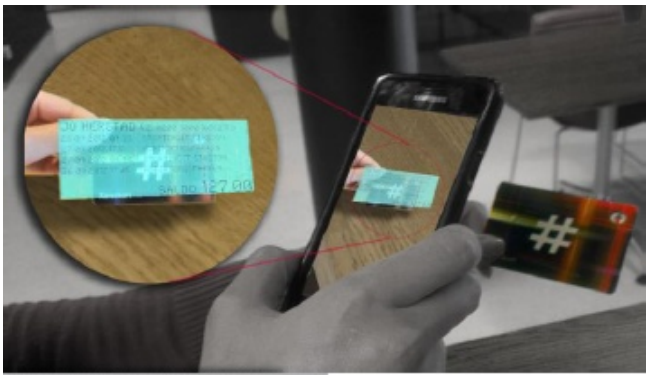


Figure 1. Giving visual immediacy property to smart tickets using an app.

A language that effectively supports visual reasoning is still not sufficiently developed [28], [29]. Even common agreement on a vocabulary of concepts that are relevant for visual thinking and reasoning is lacking. An initial vocabulary for talking about visual reasoning was offered in [30]. It included concepts such as *visual immediacy*, *impetus*, *impedance*, and *blending, analogies* and *associations*. Visual immediacy, unlike the much discussed concept of affordance [31]–[33], facilitates reasoning and does not necessarily call for action other than the reasoning itself. Impetus nudges action and impedance, in line with affordance, is responsible for instinctual negative response to visual input. In [34], these concepts were applied to web design. A website that, at a glance, enables a user to understand what the site is about and how to navigate it has visual immediacy. It has impetus if it nudges a user to engage with the site, and impedance if there are hindrances to engagement. These three design characteristics may also be used in a wide variety of sense-making situations in HCI.

Association can be defined as "*the representation of a familiar system by means of visual attributes corresponding to a different system, in order to make the user associate the two systems"* [35]. This is different from visual analogy, which is defined as "*the representation of a new system by means of visual attributes corresponding to a similar system, familiar to the user* [35]. Abduction refers to a reasoning process in which a pre-condition is inferred from a consequence, but the pre-condition is not necessarily the only one or the right one. Lastly, blending is about blending two dissimilar concepts. These concepts are perhaps easier explained using images from Absolute advertising campaigns [36], see Fig. 2.

Figure 2. Absolute Vodka ads make great use of visual association, analogy, abduction and blending.

IV. The Ecology of Experience

Service Design (SD) is a multidisciplinary field that gained momentum with the introduction of design thinking [37]–[40], where visual thinking is an important attribute of design thinking processes [41]. Service design draws often on methods familiar to HCI researchers, such as card sorting, scenarios, role-playing, personas, focus groups and observations.

The term *customer experience design* in SD is understood as a holistic concept, which integrates all aspects of a service. In other words, design for good customer experience implies good service design using user-centered design methods. The service may include several providers, but is considered as one service as long as customers experiences the service as one [42].

Service Design may also be defined in terms of experiences as a "*design for experiences that happen over time, and across different touch points"*, a definition given by Clathworthy [43]. A *touch point* is one of the central concepts in SD, together with *customer journeys*, *touch points, ecology of experiences, and service design cards*. We now define these concepts.

A. Customer Journeys

Customer journey is one of the most effective visual tools in service design. It is similar to storyboards and use cases in HCI, helping to visualize a service in an organization or a company. In [44], Koivisto explains customer journeys as follows: "*Services are processes that happen over time, and this process includes several service moments. When all service moments are connected, the customer journey is formed. The customer journey is formed both by the service provider's explicit action as well as by the customer's choices*".

We consider customer journeys to be formed not only by service moments, but also include all the experiences within and between those moments and user's responses to those experiences.

B. Touch Points

A customer journey is comprised of touch points, the service moments as described by Koivisto [44], or nodes in a visual, graphical representation of a journey. A touch point forms a link between the provider and a customer, and as such is the origin of customer experiences with the service in question. Touch points form one of the three pillars of service design [44, p. 142].

While touch points are a fundamental part of service design and a starting point in re-design of services, we consider the intervals between them to be important for user experience design.

C. Ecology of experiences

An approach to understanding experiences may be that of Nardi and O'Day [45], who use the term 'information ecology' to describe an interrelated system of people, practices, values, and technologies within a particular local environment. This ecology approach, applied to service ecology [46], and the framework for studying user experiences while interacting with technology developed by Forlizzi and Battarbee [47], shaped our theoretical perspective. *"Experiences and emotions are not singular events that unfold without a relationship to other experiences and emotions",* [47].

Building forth on these understandings, we define *ecology of experiences* as an interrelated, scalable set of experiences along a particular customer journey. In this paper, the context for creating customer journeys is that of travelling with public transportation.

D. Service Design Cards

A tool to address the touch points in the initial stages of service development is a set of service design cards, see [43]. Clathworthy provides six different use contexts for his all-purpose card set and evaluates the cards based on their intended function. The cards were found to be helpful in team-building activities in cross-functional teams. Further, they were found to be helpful in assisting with the analysis and mapping of existing situations, generating ideas for new solutions or approaches, needs elicitation and facilitation of communication.

V. Development of the Travel Experience Cards

Tangible objects, such as cards, and the images depicted on them, are known to facilitate visual reasoning and help with finding a common language for communication among people with diverse backgrounds [2], [14]. The common understandings are built through negotiation and discussion of associations and concepts related to images.

The Service Design AT-ONE cards described earlier, [43], provided the initial inspiration for the Travel Experience Cards (TEC). In this section, we will describe the design of the TEC card set for working with experiences in public transportation, and some of the ways in which the card set can be used.

A. Making the TEC Card Set

We used participatory observation and photographic documentation [48] to record our own and other travellers' experiences, collecting a large number of relevant photographic images, representing touch points and experiences while commuting, using public transportation.

All users of urban public transportation plan their trips in some way. Perhaps the starting touch point for a commute is a smart phone app, or a web-based service. The next touch point may be purchasing the ticket on the smart phone, or on the machine at the station. Digital boards may show information on trains or other means of transportation. Whatever the touch points on a particular trip are, they are part of some phase of the commute, as shown in Fig. 3. These phases are: planning the trip, making sure one has a valid ticket, arriving to a stop, embarking, traveling (this can be interrupted by, for example an accident, a ticket control, or other forms of disruption), disembarking, perhaps repeating some of the steps if transfer was needed, arriving to a final station and arriving to a final destination.

The images for the cards in Fig. 3 were not home-made, rather they were found on the net, intentionally different in style than the images we collected. A purple colored stripe was used to further differentiate these cards and formed the background to the description on the card.

Figure 3. Cards with a purple stripe represent phases of a typical trip from one destination to another.

The images collected as representations of touch points and user experiences in public transportation were then sorted into pre-determined categories corresponding to the phases of a typical commute. We initially had too many cards in each category and those images were selected that best represented the user experience. As most cards used in card sorting, ours consisted of the image and the text. Inspired by the Absolut concept [36] of using two words, our text was just one or two words long. The words were chosen for each image and typed on a red background. The first set of TEC cards was thus made, consisting of two different types of cards, those representing phases of the

customer's trip, and a mixture of cards representing touch points and experiences, see Fig. 4.

In order to ensure that images convey appropriate experiences and that the text is suitable, we have done quick-and-dirty user testing: we have simply shown the cards and asked two students (who also are the public transportation users) what they see on cards and if words match what the image conveys. At this stage, we did not want a perfect set of cards, but rather, the one that that was open for modifications and additions. For example, we chose not to make separate cards for embarking and disembarking, even though one of our testers suggested it. We wanted to see if distinctions in experiences between these two segments were important for users. If they were, separate cards would be designed for the final set.

Figure 4. Touch point cards related to the 'Station' segment, such as an e-ticket or a mobile app ticket, and experience cards, e.g., feeling safe, having an access to a convinience store, coffee, somewhere to place a bike.

B. The Initial Use Methodologies

There are two components to the TEC card set: the TEC *cards* and the TEC *use modes*. The latter are ways in which cards are used with users during testing or workshops. We have worked with two use modes, both of which used association as a way to elicit information on experiences. The first TEC use mode was based on a *forced association concept*. This use mode was employed in relation to every card representing a phase of the commute. *Focus event* was our second TEC use mode. A specific, significant event in person's life, related to the use of public transportation, was the focus of the discussion. By significant, we meant an event that is out of the ordinary, either positive or negative. For example, losing a wallet on a city bus, with driver's license and a whole lot of other important documents, would be en example of such event. The cards, both experience/touch points and journey phases, relevant for a focus event were selected from the card set, and their influence on the event discussed. The focus event was based on the same experiences as the rest of the workshop (e.g., safety, joy, being on time).

The first workshop was a pilot workshop, in which three researchers tried different use modes, and how they elicit information. For example, concern was whether working with touch points and customer journeys was better than using phases of the commute in order to understand diverse user experiences. The use mode that we agreed worked best was a forced association. The experience of *safety* in urban public transport was used as a test case, see [1] for details.

In the second workshop, forced association using the experience of *joy* in public transportation was implemented with two of the authors and three users of public transport. Two of the users were PhD candidates and the third a master student in the design, use and interaction study program. The last workshop, on the experience of *arriving on time* included the authors, and two students, one PhD and one master student in the same program. These two workshops required about an hour each time, see Fig. 5. We used convergent and divergent conversations, opening up for stories, reflections and memories, but also sense making of these, now collective, experiences, see [1].

The purpose of the workshops was to work with the use methodologies, in order to gather and understand information that could serve as the basis for designing better travel experiences. We did not study these experiences themselves.

Figure 5. A whorkshop where cards are used in conjunction with every segment of the trip, addressing just one kind of experience at the time.

In the workshops we focused on the first research question: how to design a good, context specific card-based tool. In order to address the second and the third question, how to make use of visual methodologies to discuss user experiences and which concepts support well these discussions, we have conducted two additional workshops.

C. Exploring the Visual Reasoning Framework

We engaged two professional designers with long experience with visual materials and user experience design, in an hour-long session. The aim of the session was to evaluate some of the images used for cards using all seven concepts for each image, and to discuss the findings.

The designers were asked to look at each card, first without any text, in turn, for couple of seconds. Then, they wrote down what their understanding of the card was, were there any associations with the image, any impetus, impedance, and other concepts. When done, the cards with text were shown. The cards were then discusses in the light of the intended meaning, and usefulness of concepts evaluated. The cards chosen for this purpose are shown in Fig. 6, and findings presented in the discussion section.

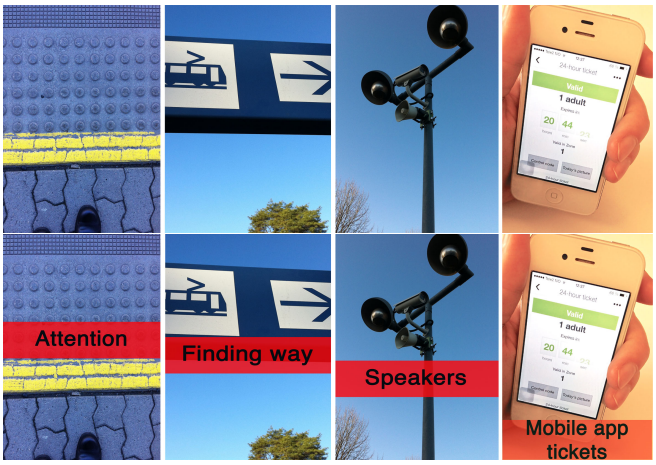


Figure 6. Images were first shown without text, then with text to two professional designers. All seven concepts were tried on just images.

Further, a workshop was organized to find out if an explicit focus on the framework could add value, e.g., generate other use modes, produce richer set of data on experiences when traveling, evaluate the quality of cards and, most importantly, how it supports knowledge production.

During the workshop we used a TEC card set consisting of 68 cards. Five participants took part: the authors and three colleagues or PhD students. The cards were placed face down on a table and the workshop participants were instructed by one of the authors. During the first hour, we used seven concepts for visual reasoning, *immediacy, association, analogy, impetus, impedance, abduction, and blending,* to talk about the cards. Thus the question the participants needed to answer after looking at a card was: "Does this card enables … (*immediacy, association, analogy, impetus, impedance, abduction, and blending*) (see Fig. 7). The participants had 10 seconds to describe the first concept, *immediacy*, and 30 seconds for the other six concepts. Each participant wrote down the answers on a piece of paper and the results were discussed in the group. We started with each participant taking three cards from the pile and describing each card using the *immediacy* concept.

This was repeated with the concepts *association, analogy, impetus, and impedance*. Each participant used only one card for the last two concepts, *abduction and blending*. During the second part of the workshop, each participant took one card only, but applied all seven concepts to this card, registering their findings, similar as in the session with professional designers. This was followed by the discussion of findings.

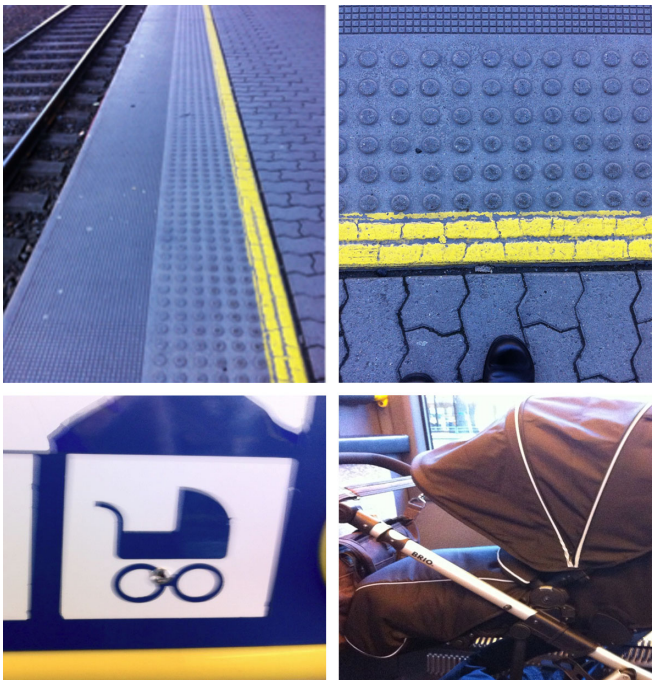


Figure 7. Which of the images provokes more associations, allows for abduction and blending, minimizes impedance?

VI. Discussion

All four workshops had a small number of participants, the smallest one just 3 and the largest 5. However, all participants had a solid background in both user experience design and various methods of working with users, including co-design, participatory design, and user-centered design. This is relevant because most of them have worked with similar methods before and could give qualified opinions about the use methodologies. All participants were at the same time also users of public transportation. We felt that workshops with a small number of participants worked well at this phase of the project. Participants with such background provided good feedback on the TEC Card Set, the use modes (the forced association and the focus event), and work with the visual reasoning framework.

A. Working with images and visual reasoning concepts

Diverse insights were gained from the user session with professional designers. In the discussion after working with concepts, the two participants both said that visual immediacy was very interesting, and should be further explored. At the same time, their interpretation matched only 75% intended meanings of images that they worked

with. For example, one participant interpreted the image of laud speakers as a surveillance camera, Fig. 6, and the other interpreted the mobile app ticket as a generic app. When the text was added, the interpretations became clear. The participants, however, suggested that a shorter text reading simply 'ticket' or 'mobile ticket' should replace 'mobile app tickets'.

Two images were seen as giving impetus to action, 'press the button' and 'turn right for the train'. There were no hindrances to understanding images perceived, even though, as mentioned, the images were not always interpreted correctly. All images allowed for associations, positive or negative, as well as analogies (with street lights, other apps, separation, big cities and others). Abduction was needed in conjunction with a mobile app ticket, as well as with the image of a yellow line that signifies attention. The participants felt that blending was not represented in any of chosen images, and that the use of two or three provided images together, was also difficult.

B. *Vocabulary*

It was important for the participants to understand the TEC cards. Only then could they really engage in working creatively with them. As it was not possible to have a card representing each individual experience, the terms describing the cards were chosen with care. We found out that some cards needed to be broad enough to allow for several different interpretations. Others, as for example *ticket* needed further specification: *valid ticket* and *price of ticket* were the requests from our participants. There was also a suggestion to further specify attributes relevant to the validity of the ticket, such as the visibility of information.

C. *On the use the cards in workshops*

The workshops with our participants started with explanation of the purpose of the workshop, the TEC set, and how we were going to use the data in the future. We then asked the participants to focus on what gives them, as users of public transportation, the experience of *joy* or *arriving on time*. During the workshop on joy, it became clear that we were missing several experience cards: weather, space, valid ticket, toilet, charging battery, time, and price of ticket. During the workshop on arriving on time, experiences that help users or are a hindrance to reaching their final destination on time were considered. During this exercise two new experience cards were proposed: *ticket control* and *event*. Ticket control was perceived as both a segment card and an experience card, experience of control of the ticket validity. The event experience card refers to large events, such as sports championships and matches, in which large crowds of people use public transportation. During these events it is often impossible to arrive on time.

By constructing common understanding and meaning giving to the cards, for the entire length of the trip, we found that a number of combinations of experiences have emerged as important. For example, a card with a term 'crowd' was used extensively. It was related to several segments (station, ticket, embarking and disembarking and traveling) to both feeling of lack of safety, lack of joy and danger of being late. One of the participants then mentioned that there is really nothing one can do with this knowledge. This started a whole discussion on the strategies that people use to avoid crowds. At the end of the discussion, all participants agreed that, actually, there are opportunities for making things better by design.

The same conclusion was reached regarding the use of cards to address focus events. We illustrate this with two examples. One participant told a story of a woman who had a very unpleasant experience on the train. She never enjoyed taking public transportation again, and never took trains very early in the morning or late in the night. The card that she held while talking about the experience depicted a station in dusk, empty, and not giving the feeling of being safe (see Fig. 2).

The second example had to do with embarrassment over being caught without valid ticket and blaming several touch points involving technology that were not working properly at the time. In both cases, cards representing related experiences were found and participants considered frequency of such events, their impact on people's lives, possible design solutions, etc.

The seven concepts were very useful in evaluating the cards, because they invite a particular kind of visual reasoning that is more comprehensive than discussing the meaning of the card. The design of the TEC Card Set is as such that the keyword printed on the card provides the meaning of the card in cases where the image is unclear or is perceived as for decorative only. The visual reasoning framework focused us on the image; each concept became a lens that both explored and mediated the meaning of the image.

Figure 8. The card on the left uses green to point to the emergency exit. To some people, green is not a colour that infers emergency. In the picture to the right, it is not possible to infer anything about the conductor.

Using the framework we found some cards that are strong candidate for replacement, such as the image in Fig. 8, on the right. On the other hand, the quality of a card was not based on a direct translation between the image and the keyword, but on how the image itself triggered visual reasoning. The combination image and concept was crucial here. We did not evaluate the seven possible combination of each card, but our initial findings do encourage us to continue this line of thinking in the next step of our research.

The main challenge will be how to weigh each concept. For example, is immediacy a more important concept than association? The concepts immediacy and association were easy to engage with. Visual immediacy is an important characteristic of a design card in general, as it is about first impression and meaning making, which may have a direct affect on further work with the cards. Association plays a similar role, but is based on interference from one conceptual domain to another. Association focuses us on what is possible and thus broaden the field of possible positive and negative situations and experiences. The best use modes we found so far, were building on associations.

D. User's comments

After each workshop, a few minutes were set aside for asking the participants about their experiences of working with the TEC cards. One of the participants (male, 39) said: *They were good to get the conversation going and explore different topics in a quick and easy manner.* Another participant told us: *The images put you kind of into a memory lane. When I look at the station card, I remember my own station and I can feel the experiences. They make me more aware of the things I should think of. I would never come up with as many examples of experiences as we jointly did* (female, 27). Asked whether it was boring to repeat forced association technique, the participants agreed that it was a good experience, connecting the detailed pictures around each segment card into a larger picture, which was more relevant: *This was actually a learning experience for me. Cards with good quality images and nice colours made association easier* (female, 26).

E. Reasoning with visual concepts

We found that the concepts *immediacy* and *association* were easy to use. All participants could use these concepts to talk about the cards as well as to give feedback on the quality of the cards. For example, the card showing a nice cup of coffee had high visual immediacy, while a card portraying a recycling trashcan had low immediacy. Evaluation of the cards used for immediacy showed that *the number of objects on a card and the organisation of these objects affected visual immediacy*. A*ssociation* was a good concept to bring out different understandings of the quality of images, and colours in particular.

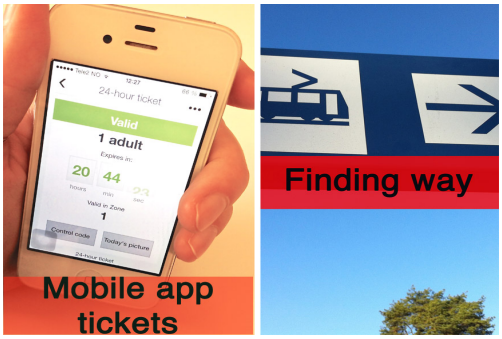


Figure 9. Cards with good quality images and nice colours made association easier.

For example, if the colours were bright and beautiful (Fig. 9), participants found it easier to make associations and these associations tended to be the positive ones. We found out that in general, all cards used for association were easy to use for this purpose.

The *impetus and impedance* concepts were helpful in establishing if the general meaning of a card was communicated well. On the other hand, we found that some of the cards seem too abstract to support the concepts *analogy and abduction*. Working with these concepts resulted in a focus on particular aspects of a card, such as colour. Participants established that the use of red and green was often problematic, as some of the people assigned the meaning to these colours based on the traffic light analogy, while others considered the actual context. Some cards did not have sufficient visual information in order to infer (use abduction) to understand the meaning, see Figure 8, the 'conductor' card.

Conceptual *blending* was a creative way to work with the cards, as it enabled a better understanding of the context. For example, blending *immediacy* with *impetus* was a good match and easy to explain, e.g., 'what is the first thing you want to do when you see this card'? Since we used only four cards for blending, we did not explore its potential fully. But we could see the indications that it would be beneficial for design of new experiences, as it enables discussions over more complex domains, visually.

VII. Conclusion and Future Work

Understanding user experience is important in the design of interactive products and services. People's experiences with public transportation, even with a single touch point, such as ticket validation, are very different. This heterogeneity makes working with user experiences challenging.

The TEC set was found to respond to this challenge adequately. Heterogeneity remained visible, yet a common understanding of an experience emerged during the workshops, when working with safety, joy and being on time.

The size and the feel of cards were found to be satisfactory. Part of their appeal was attributed to the images. The images were taken out in the field, thus from users actual context, but were generic enough to easily evoke memories of many diverse experiences. The other part of the appeal was tangibility of the cards. They served as tangible pointers to experiences, evoking memories and facilitating conversation about experiences. They enabled rich communication, in depth when working with focus events, and in breadth when working with forced associations across all segments of a customer journey. Our focus was not on re-designing services at this time, yet many ideas and thoughts that emerged on during the workshops would be worth pursuing further. The number of cards could be reduced. At times, it was overwhelming to search the set of 68 different cards. In later work, we have used about half of that number.

While these conclusions are in line with previously published work and thus not new, we hope that we have explained the process of creating the tool and its evaluation

(by users of public transportation, with good understanding of design processes) in such a manner that it is inspirational. This methodology of creating a set of cards for studying user experiences is rather fast and fun. The set can be used to understand a range of experiences in a given context of use, both in breadth and in depth, identifying clear design and innovation opportunities.

We have tested a small number of TEC use modes (forced association and focus event). While doing that, new possibilities based on the framework concepts have opened up, and future work will explore them further. *Visual immediacy* and *conceptual blending* appear to be two strong candidates as the bases for new use modes.

Finally, we brought this visual methodology further into other context specific application areas, such as service innovation and customer experiences in the library and the experiences related to transition of young patients from children's hospitals to adult ones. The card sets developed for these purposes have worked very well in sense-making, exploratory phases of design processes and helped shape further design efforts.

REFERENCES

[1] A. L. Culén, M. van der Velden, and J. Herstad, "Travel Experience Cards: Capturing User Experiences in Public Transportation," in *The Seventh International Conference on Advances in Computer-Human Interactions*, 2014, pp. 72 - 78.

[2] K. Halskov and P. Dalsgaard, "Inspiration Card Workshops," in *Proceedings of the 6th Conference on Designing Interactive Systems*, New York, NY, USA, 2006, pp. 2–11.

[3] J. Brucker, "Playing with a Bad Deck: The Caveats of Card Sorting as a Web Site Redesign Tool," *Journal of Hospital Librarianship*, vol. 10, no. 1, pp. 41–53, Jan. 2010.

[4] S. Pink, "Interdisciplinary agendas in visual research: re-situating visual anthropology," *Visual Studies*, vol. 18, no. 2, pp. 179–192, Oct. 2003.

[5] W. Hudson, "Playing Your Cards Right: Getting the Most from Card Sorting for Navigation Design," *interactions*, vol. 12, no. 5, pp. 56–58, Sep. 2005.

[6] "The World leader in Card Sorting Tools | Optimal Workshop." [Online]. Available: https://www.optimalworkshop.com/optimalsort. [Accessed: 10-Mar-2015].

[7] C. Wölfel and T. Merritt, "Method Card Design Dimensions: A Survey of Card-Based Design Tools," in *Human-Computer Interaction – INTERACT 2013*, P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, Eds. Springer Berlin Heidelberg, 2013, pp. 479–486.

[8] A. L. Culén and A. Gasparini, "Find a Book! Unpacking Customer Journeys at Academic Library," in *The Seventh International Conference on Advances in Computer-Human Interactions*, 2014, pp. 89–95.

[9] A. L. Culén, "Later Life: Living Alone, Social Connectedness and ICT," in *6th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*, 2015.

[10] "Method Cards | IDEO." [Online]. Available: http://www.ideo.com/work/method-cards/. [Accessed: 10-Mar-2015].

[11] "PLEX CARDS · Playful Experiences Cards." [Online]. Available: http://www.funkydesignspaces.com/plex/. [Accessed: 10-Mar-2015].

[12] "Design with Intent | Design patterns and human behaviour." [Online]. Available: http://designwithintent.co.uk/. [Accessed: 12-Mar-2015].

[13] S. Clatworthy, "The AT-ONE touch-point cards [Online]. Available: https://www.academia.edu/7288723/The_AT-ONE_touch-point_cards_for_printing. [Accessed: 10-Mar-2015].

[14] D. Spencer, *Card Sorting: Designing Usable Categories*, 1st edition. Brooklyn, NY: Rosenfeld Media, 2009.

[15] S. Pink, *Doing Visual Ethnography*. SAGE Publications, 2007.

[16] S. Pink, *Advances in visual methodology*. Sage, 2012.

[17] G. Rose, *Visual Methodologies: An Introduction to Researching with Visual Materials*. SAGE, 2012.

[18] S. Lindsay, D. Jackson, G. Schofield, and P. Olivier, "Engaging older people using participatory design," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, New York, NY, USA, 2012, pp. 1199–1208.

[19] "PhotoVoice : About PhotoVoice." [Online]. Available: http://www.photovoice.org/about/. [Accessed: 10-Sep-2014].

[20] H. Ishii and M. Kobayashi, "ClearBoard: A Seamless Medium for Shared Drawing and Conversation with Eye Contact," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 1992, pp. 525–532.

[21] A. L. Culén, H. N. Mainsah, and S. Finken, "Design Practice in Human Computer Interaction Design Education," in *The Seventh International Conference on Advances in Computer-Human Interactions*, 2014, pp. 300–306.

[22] S. Finken, A. L. Culen, and A. A. Gasparini, "Nurturing Creativity: Assemblages in HCI Design Practices," in *Proceedings of DRS 2014*, Umeå, 2014, pp. 1204–1217.

[23] S. Bødker, "When second wave HCI meets third wave challenges," in *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, New York, NY, USA, 2006, pp. 1–8.

[24] S. Harrison, D. Tatar, and P. Sengers, "The three paradigms of HCI," in *Alt. Chi. Session at the SIGCHI Conference on Human Factors in Computing Systems,* San Jose, California, USA, 2007, pp. 1–18.

[25] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999.

[26] M. Hegarty, "The Cognitive Science of Visual-Spatial Displays: Implications for Design," *Topics in Cognitive Science*, vol. 3, no. 3, pp. 446–474, Jul. 2011.

[27] M. van der Velden and A. L. Culen, "Information Visibility in Public Transportation Smart Card Ticket Systems," *International Journal On Advances in Networks and Services*, vol. 6, no. 3 and 4, pp. 188–197, Dec. 2013.

[28] R. E. Horn, *Visual Language: Global Communication for the 21st Century*, 1 edition. Bainbridge Island, Wash: MacroVU Press, 1998.

[29] E. R. Tufte, *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, Conn: Graphics Press, 1997.

[30] A. Karabeg and N. Akkøk, "Towards a language for Talking about Visual and Spatial Reasoning," in *Visual Literacy And Development: An African Experience*, R. Griffin, S. Chandler, and Cowden, Belle Doyle, Eds. The International Visual Literacy Association, 2005, pp. 109 – 115.

[31] V. Kaptelinin and B. Nardi, "Affordances in HCI: Toward a Mediated Action Perspective," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2012, pp. 967–976.

[32] D. A. Norman, "Affordance, conventions, and design," *Interactions*, vol. 6, no. 3, pp. 38–43, 1999.

[33] J. J. Gibson, "The concept of affordances," *Perceiving, acting, and knowing*, pp. 67–82, 1977.

[34] A. Karabeg and N. Akkøk, "Visual Representations and the Web," in *Visual Literacy And Development: An African Experience*, R. Griffin, S. Chandler, and Cowden, Belle Doyle, Eds. The International Visual Literacy Association, 2005, pp. 115–123.

[35] A. Karabeg, M. N. Akkok, and K. Kristensen, "Towards a language for talking about information visualization aimed at presentation on the Web," in *Eighth International Conference on Information Visualisation, 2004. IV 2004. Proceedings*, 2004, pp. 930 – 937.

[36] "Absolut Ad." [Online]. Available: http://www.absolutad.com/. [Accessed: 10-Sep-2014].

[37] R. L. Martin, *Design of business: Why design thinking is the next competitive advantage*. Harvard Business Press, 2009.

[38] T. Brown, "Design Thinking," *Harvard Business Review*, vol. 86, no. 6, p. 84, 2008.

[39] T. Brown, *Change by design: how design thinking can transform organizations and inspire innovation*. New York, NY: Harper Collins Publishers, 2009.

[40] A. Culén and M. Kriger, "HCI in IT-facilitated Business Innovation: a Design Thinking Perspective," presented at the 16th International Conference on Human-Computer Interaction, Crete, 2014. *HCI in* Business *Lecture Notes in Computer Science,* pp. 492 - 503.

[41] G. Goldschmidt, "On visual design thinking: the vis kids of architecture," *Design Studies*, vol. 15, no. 2, pp. 158–174, Apr. 1994.

[42] L. G. Zomerdijk and C. A. Voss, "Service Design for Experience-Centric Services," *Journal of Service Research*, vol. 13, no. 1, pp. 67–82, Feb. 2010.

[43] S. Clatworthy, "Service innovation through touch-points: Development of an innovation toolkit for the first stages of new service development," *International Journal of Design*, vol. 5, no. 2, pp. 15–28, 2011.

[44] S. Miettinen and M. Koivisto, *Designing Services with Innovative Methods*. University of Art and Design Helsinki, 2009.

[45] B. A. Nardi and V. L. O'Day, *Information ecologies: using technology with heart*. Cambridge, Mass.: MIT Press, 1999.

[46] A. Polaine, L. Løvlie, and B. Reason, *Service design: From insight to implementation*. Brooklyn, N.Y.: Rosenfeld Media, 2013.

[47] J. Forlizzi and K. Battarbee, "Understanding experience in interactive systems," in *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, 2004, pp. 261–268.

[48] M. Crang and I. Cook, *Doing ethnographies*. Los Angeles; London: SAGE, 2007.

Spatio-Temporal Density Mapping for Spatially Extended Dynamic Phenomena

- a Novel Approach to Incorporate Movements in Density Maps

Stefan Peters
Department of Geoinformation
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
e-mail: stefan.peters@directbox.com

Liqiu Meng
Department of Cartography
Technische Universität München
München, Germany
e-mail: liqiu.meng@bv.tum.de

Abstract - **The visualization of density information and its changes is a crucial support for the spatio-temporal analysis of dynamic phenomena. Existing density map approaches mainly apply to datasets with two different moments of time and thus do not provide adequate solutions for density mapping of dynamic points belonging to the same moving phenomena. The proposed approach termed as Spatio-Temporal Density Mapping intends to fill this research gap by incorporating and visualizing the temporal change of a point cluster in a 2D density map. At first either straight or curved movement trajectories based on centroids of spatio-temporal point clusters are detected. The traditional Kernel density contour surface is then divided into temporal segments, which are visually distinguished from one another by means of a rainbow color scheme. Furthermore, several ideas for the investigation of the usability of our approach are addressed.**

Keywords - spatio temporal density map; rainbow color scheme; visual analytics; dynamic phenomena.

I. INTRODUCTION

As stated in a previous work by Peters and Meng [1], visualization helps to investigate and understand complex relationships in a spatial context. Maps account as one of the most powerful visualization forms. They represent geographic information in abstract ways that support the identification of spatial patterns and the interpretation of spatial phenomena. Furthermore, the visual presentation and analysis of dynamic data and dynamic phenomena is currently a hot research topic [2].

Hence, in today's society, the need for data abstraction along with the growing amount of available digital geodata is rapidly increasing. One reasonable way of abstracting data is provided by density maps [3]. Density maps can be applied for point data in various fields, for instance, in physical or human geography, geology, medicine, economy or biology [4, 5]. How to present the density for dynamic data/phenomena is, however, not yet adequately addressed.

In this paper, we introduce a novel density mapping approach for spatially and temporally changing data. The approach is based on [1], whereby in this work a different test dataset was used, two different types of movement trajectory concepts are introduced, a verification of the used rainbow color scheme is presented, investigations about wrongly assigned points are considered and a more detailed comparison between STDmap and its alternative in form of Kernel Density Estimation (KDE) maps for each temporal interval is provided.

In the next section, the state of the art related to density maps, in particular, an overview of approaches considering the dynamics of movement data in the density visualization is given. In the section afterwards, our own approach is described in detail, followed by implementation processes, discussions of the results and a conclusion.

II. DENSITY MAPS - STATE OF THE ART

One of the most straight forward ways to visualize point density is a scatter plot or a dot map. Graphic variables for point symbols, such as size, shape, color and transparency, can be applied in relation with the attribute value. In order to discern the density distribution, these graphic variables can be iteratively adapted to the given map scale, but still the occlusion of neighboring points cannot be always desirably avoided. The density value of each point can be obtained by counting all points within a buffer around the point or within a grid cell the point is located in.

In the following, the density estimation and map principles are shortly presented and the state of the art of density maps with static or dynamic data is given.

A. KDE

KDE [6] is a classic method widely used to determine densities of individual points that represent a continuous surface. The KDE approach is described in detail in [6-8]. The standard KDE, a normal distribution function, uses a Gaussian kernel:

$$\hat{f}_h(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K(u) \qquad (1)$$

whereby $u = \frac{X - X_i}{h}$ and $K_G(u) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}u^2\right)$

with: $\hat{f}_h(x)$ = general Kernel density function
K = Kernel function
K_G = standard Gaussian function
h = smoothing parameter (bandwidth)
n = number of points
X = point (x,y) for which the density will be estimated
$X_1, X_2, ..., X_N$ = sample points, placed within the kernel radius h

Beside a Gaussian kernel, also other kernel types can be applied, such as Triangular, Biweight, Epanechnikov or Uniform kernel [7]. A certain bandwidth (search radius) is defined for the kernels, located around each point. For each cell of an underlying grid (defined by a certain resolution) a density value is calculated as shown in equation (1) and hence a smooth surface is provided [9]. The kernel bandwidth value strongly effects the resulting density surface [10]. A formula for an optimal bandwidth is offered by Silverman [7] as shown in equation (2).

$$bw\_optimal = 1.06 * \min\left(\sqrt{\mathrm{var}(P)}, \frac{\mathrm{IQR}(P)}{1.34}\right) * n^{-\frac{1}{5}} \quad (2)$$

with: *bw_optimal* = optimal bandwidth
P = point dataset (coordinates)
IQR*(P)* = interquartile range
var*(P)* = variance of *P*
n = number of points

In order to detect clusters, KDE has been applied in in various applications, such as crime analysis and population analysis. Kwan [11] used KDE and 3D visualization to investigate spatio-temporal human activity patterns. The author applied the density estimation as a method of geovisualization to find patterns in human activities related to other social attributes. The classic KDE was investigated in [12-14], and thereby defined as a visual clustering method. In these works, KDE maps were created in order to visually provide a better overview and insight into the given data.

B. Contour lines and intervals

A common technique to map point densities calculated using KDE are isopleth maps with filled contour intervals. The term "isopleth map" refers to one of two types of isoline maps (also called isarithmic or contour maps). In the first type of isoline maps each contour line indicates a constant rate or ratio derived from the values of a buffer zone or kernel area. In this sense, the continuous density surface is derived from an originally discrete surface. In the other type of isoline maps (commonly referred to "isometric map"), contour lines (isometers) are drawn through points with directly measurable equal value or intensity such as terrain height or temperature [15]. It is assumed that the data collected for enumeration units are part of a smooth, inherently continuous phenomenon [16]. In our context, we only use contour lines to delimit the intervals (the areas between contour lines).

Furthermore, Langford and Unwin [17] provided a good overview of density surfaces used in Geographic Information Systems (GIS) as choropleth population density maps, population density on grids, population density surfaces, and pseudo-3D population density surfaces. In several works as [4, 5], the KDE concept is adapted for the 3D space density mapping of static 3D data.

C. Dynamic data and density information

In the following sections, an overview is given about existing works related to density maps of dynamic points.

1) Sequence of KDE maps for dynamic points

A straightforward way of visualizing the density of dynamic points would be a sequence of density surfaces (one per time interval). The change of the density in time could be better discernable by means of an animation of these density maps. We could also arrange the local density contours of each time interval on the same map. Transparency and a unique color scheme for each time interval could be applied in order to distinguish different density contours. However, the tinted intervals may spatially overlap and make the map reading a difficult endeavor.

2) Dual KDE

Jansenberger and Staufer-Steinocher [18] analyzed two different point datasets recorded within the same area, but at two different moments of time. The authors suggest a Dual-KDE approach, which results in a map illustrating the spatio-temporal density difference of the two datasets. The absolute difference is used, that is, the absolute density of the second point dataset subtracted from that of the first one.

3) DKDE

The approach called Directed Kernel Density Estimation (DKDE) that takes the dynamics of moving points inside density maps into account was suggested in previous works [19-24]. The DKDE is applicable for discrete moving points and it considers two moments of time. Instead of an upright kernel as in the KDE method, a tilted kernel is used, as illustrated in Figure 1.

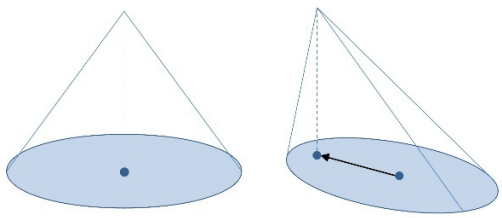


Figure 1. Linear kernel (left) and Directed linear kernel taking point speed and movement direction into account (right), source: [24].

The tilt depends on the movement direction vector of the respective point. The resulting DKDE-map shows the so-called "ripples", which can be interpreted as an indicator for the movement direction and density change of points that are located closely to each other with very similar movement speeds and directions. These ripples are visible among overlapped contour lines. The tinted contour intervals do not contain the information about movement or density change.

Peters and Krisp [24], for instance, used 2D airplane positions in the area of Germany at two moments of time with a time lag of five minutes. The resulting DKDE-map is shown in Figure 2.

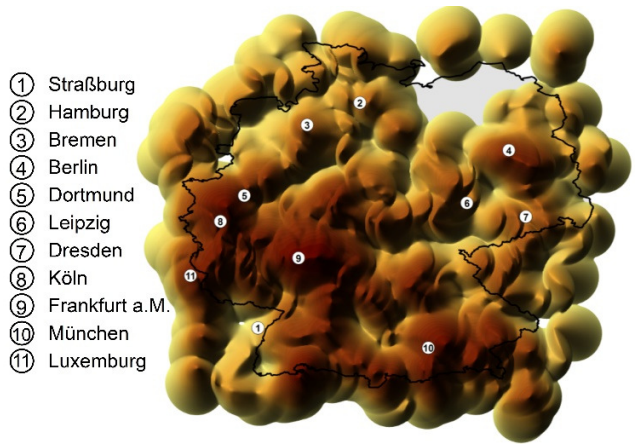


Figure 2. DKDE-map based on airplane positions at two moments of time, source: [24].

4) 3D density map using space time cube (STC)

Nakaya and Yano [25] suggest a method using a STC to visually explore the spatio-temporal density distribution of crime data in an interactive 3D GIS. Thereby the author adapted the KDE by using space-time variants and scan statistics. In order to investigate the dynamics and density change, an interactive use within a 3D environment is essential.

5) KDE for trajectories

In a comprehensive review, Andrienko and Andrienko [2] discussed existing visual analysis methods, tools and concepts for moving objects. A section is dedicated to continuous density surfaces (fields) derived from trajectories or from point-related attributes. Density maps of moving objects were created on the basis of aggregated points of trajectories. A trajectory is understood as a function of time or a path left by a moving object in space. Moving objects can be confined within a network (such as cars along streets of a traffic network) or float freely over a region (boats) or in space (airplanes). Spatio-temporal density maps of trajectories were investigated in [26-29]. In these approaches, the KDE method is adapted to trajectories as a function of changing velocity and direction. Willems et al. [30], for example, built his kernels assuming constant speeds. Furthermore, McArdle et al. [31] investigated computer mouse trajectories. Thereby density maps are generated based on movement activity. For each scale, the density map is recalibrated in order to highlight the most important areas, in terms of mouse movement activity. Other approaches assume constant acceleration. The resulting density maps can reveal simultaneously large-scope patterns and fine features of the trajectories. This mapping idea was extended to the 3D space in [27], where the trajectory densities are visualized inside a STC.

Another possibility of displaying density information of trajectories is to use derived discrete grid cells, whereby each cell color refers to the amount of trajectories passing through the cell [32, 33].

D. Research questions

In the existing 2D density maps based on KDE, the time is either frozen on a certain moment or confined within a certain time interval. Consequently, the resulting contour lines do not carry information of temporal changes. Although various approaches for density visualization of trajectories have been investigated, an appropriate method for 2D density maps of moving point clouds is still missing. Whether the dynamics of spatially extended phenomena (SEPs) - represented by points – can be adequately expressed in a single contour map remains an open question. To tackle this question, we develop an approach termed as Spatio-Temporal Density Mapping or STDmapping.

III. TEST DATASET

We used lightning points recorded by LINET, a lightning detection network [34], as the test dataset. It contains altogether 8184 detected lightning in the region between Munich, Germany and Prague, Czech Republic (47°N–50°N Latitude and 11°E–15°E Longitude) on April 26th 2013 between 2pm and 7pm. Each point is encoded with its geographic coordinates (longitude, latitude) as well as the exact lightning occurrence time. The recorded height information is not considered within our approach.

Figure 3 illustrates the lightning points in form of blue dots projected onto a plane surface. The background base map contains topographic data of the area out of Open Street Map dataset. The use of such static plot of lightning data is limited for the investigation of the dynamics in lightning data. To enhance clarity of the approach, only points of the three largest lightning tracks are considered in the STD density mapping approach.

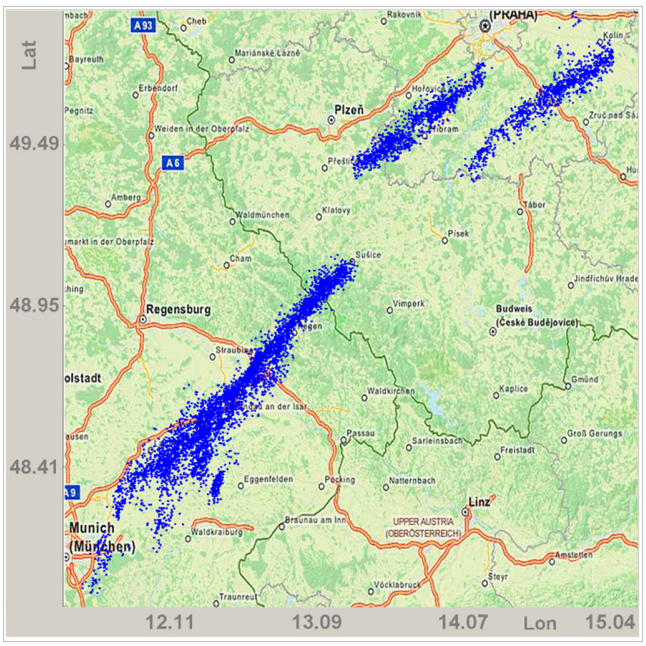


Figure 3. Initial point test dataset.

Visual analysis methods for these lightning points, which represent the moving phenomena of a thunderstorm, were published in [35-37].

IV. METHODOLOGY

First of all, density contour maps can be derived from point datasets using KDE while an optimal kernel bandwidth can be calculated according to Silverman's formula [7]. In our work we deal with a lightning point dataset representing a dynamic phenomenon. Thus, instead of creating a single density contour map of the entire point dataset, we applied KDE in each case to all points belonging to the same lightning track. Thereby, lightning points are clustered, afterwards allocated and aggregated to trajectories. In doing so, a temporal clustering is applied to the initial point dataset using a time interval of one hour. Subsequently, all points within each temporal interval are spatially clustered using a buffer threshold of six kilometers. In the resulting spatio-temporal clusters, the spatially overlapping parts within two time sequences are detected and afterwards allocated and aggregated to lightning trajectories. Further details of the temporal and spatial clustering of lightning data including explanations for thresholds can be found in [35, 36]. The results of the density contour maps derived for the test dataset are shown in Figure 4. The importance of grouping dynamic points into tracks is discussed in the next section.

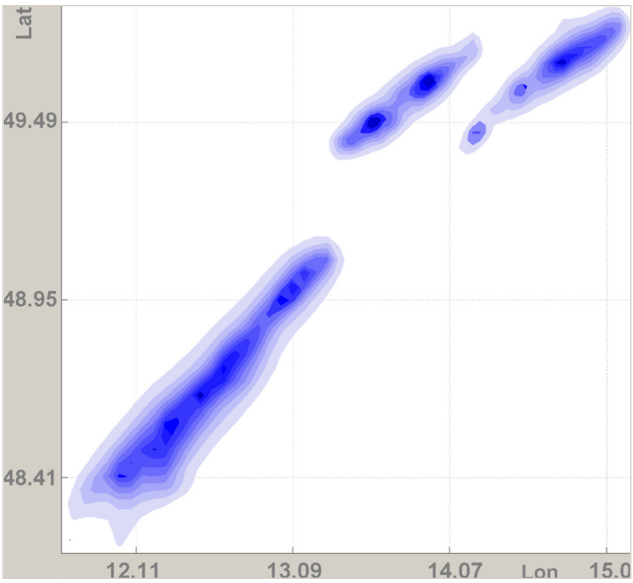


Figure 4. KDE map applied to test point dataset.

The resulting density contour layers in blue tones do not bear any temporal information. Nevertheless, the aforementioned temporal point clustering method provides time information for each lightning point. Figure 5 illustrates our initial point dataset, whereby lightning points were segmented and colored according to the different time intervals, thus reveal the dynamic changes. In doing so, we used a time interval of 1 hour starting at 2pm for the temporal clustering. The overlapping convex hulls surrounding all spatially clustered points of the same temporal interval are allocated to altogether three different trajectories.

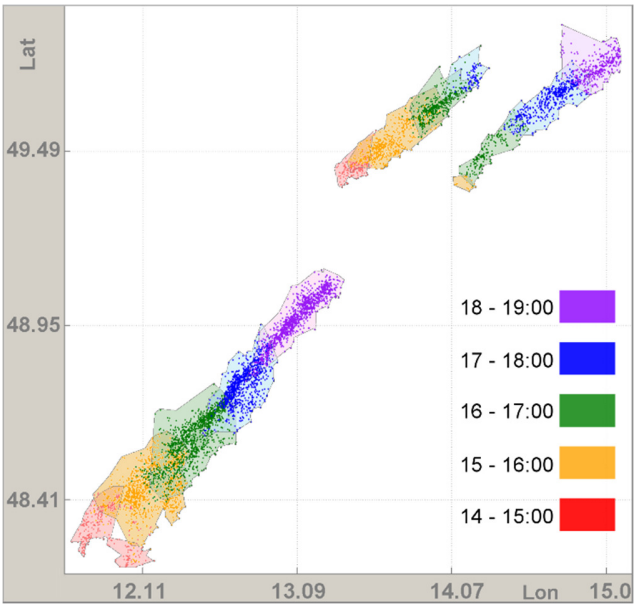


Figure 5. Temporally clustered point data.

Three main moving lightning clusters are perceivable within the test area. Their geographic and temporal locations are apart from each other with one formed at lower left part and one upper left, both starting around 2pm and the third one upper right occurring around 4pm. All clusters are moving north-eastwards. The upper left cluster disappears around 5pm, whereas the lower left and the upper right last until 7pm.

As mentioned before, traditional density mapping does not contain temporal information. Clustering and allocating dynamic point data (in our case lightning points) towards trajectories provides information about data movement (speed, direction, etc.).

In the following, we introduce a method, which includes movement information, i.e., dynamics in KDE mapping. In other words, we suggest a solution to incorporate temporal information of moving points (as illustrated in Figure 5) inside the density contour intervals (as shown in Figure 4).

A. STDmapping workflow

An overview of our suggested method is illustrated through an overall workflow in Figure 6.

First of all a density contour map using KDE is created. Additionally, the given point dataset is temporally and spatially clustered. In the next step, the overlapping clusters (in case they are temporally successive) are detected and after that allocated and aggregated to independent tracks. Cluster centroids are embedded in the trajectories. A detailed description about these steps can be found in [35].

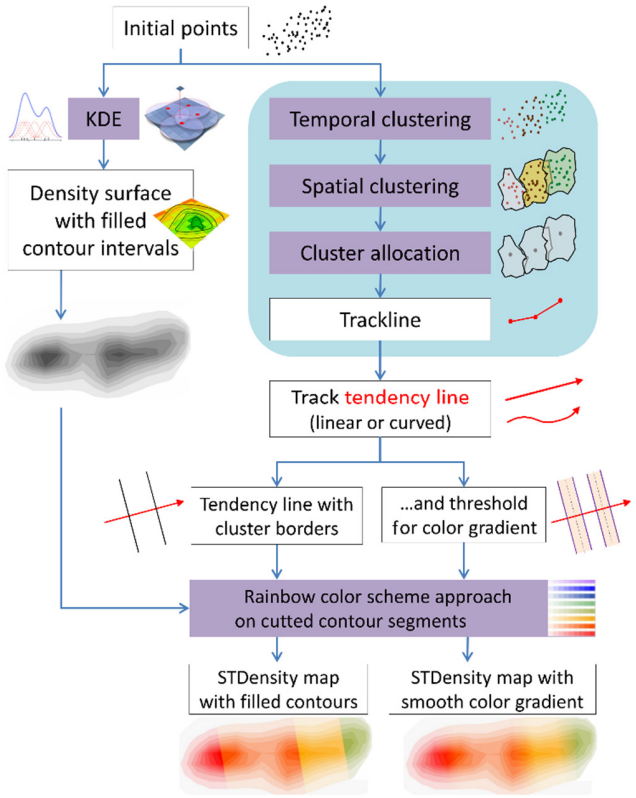


Figure 6. The workflow of STDmapping of lightning data.

A linear approximation of each track results in a straight tendency line, which represents the average moving direction of the point cluster. The linear approximation can be based either only on the cluster centroids or on the entire point datasets of a track.

If the projected trajectory is curved rather than straight, the tendency line can be approximated by a polyline connecting the cluster centroids. In our case, a cubic spline interpolation function is used to fit a curve through the cluster centroids [38].

Consequently, we have on the one hand the density surfaces represented by layered tints between neighboring contour lines and on the other hand the tendency line with either abrupt or smooth transition at borders of temporal clusters. This temporal border is a line perpendicular to the tendency line passing through the average locations of all points within a certain period (in our case 10 minutes) before and after a temporal border (e.g., full hour). If the phenomenon is moving, all points between two temporal borders (e.g., between "2pm line" and the "3pm line") are grouped into the same temporal interval (period: "2pm - 3pm").

The next question is how we can incorporate the dynamics inside the density map. The idea is to divide the tendency line into temporal parts, which will in turn guide the segmentation of the density surface.

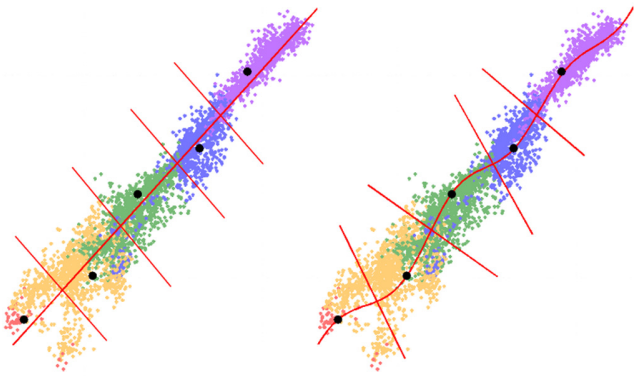


Figure 7. Temporally clustered points and cluster centroids in black with straight tendency line and perpendicular temporal borders (left) and with curved tendency line and perpendicular temporal borders (right).

Figure 7 illustrates two different ways of tendency line determination. In the left part, all points of an exemplary lightning track are colored according to the temporal cluster they respectively belong to. The cluster centroids are presented as black dots. A straight tendency line representing the general movement direction of the lightning cluster is based on the coordinates of all cluster centroids. The temporal borders in red are detected and vertically aligned to the straight tendency line. The locations for temporal borders can be defined by the half distance between two temporally successive cluster centroids, or, by the centroid of the overlapping area of two sequential temporal point sets.

In the right part of Figure 7, the tendency line is represented by a curve determined through cubic spline interpolation of all cluster centroids. The temporal borders in red are defined again as perpendicular lines of the curved tendency line. Thus, temporal border lines are not arranged parallel to each other as in the case for the straight tendency line. However, for very small temporal clusters (clusters of low velocity or very small temporal thresholds) temporal border lines are much closer to each other, and thus – due to the curved tendency line route – they are almost parallel to each other.

Nevertheless, it could be also possible that two sequential temporal borderlines intersect each other (in particular if the tendency line is strongly curved). In this case, the respective intersecting temporal borderlines need to be partly merged as shown in Figure 8. Thereby, different temporal segments are illustrated in different colors (beige, green, blue) and temporal borders in red. We suggest combining the two intersecting temporal border lines from the intersection point onwards towards the outer cluster extension in a way that each of both temporal border lines forms the same angle with the continuing merged border line part.

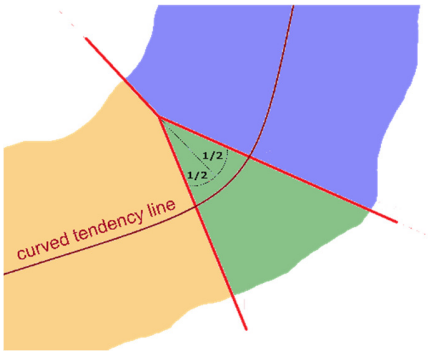


Figure 8. Suggested solution for intersecting temporal border lines (red).

In the next step, density contours are separated through temporal borderlines into temporal surface segments as illustrated in Figure 9.

As described before, temporal borders can be either parallel if they are based on a straight tendency line (see Figure 9 left) or temporal borders are perpendicular to the curved tendency line and thus non-parallel (see Figure 9 right).

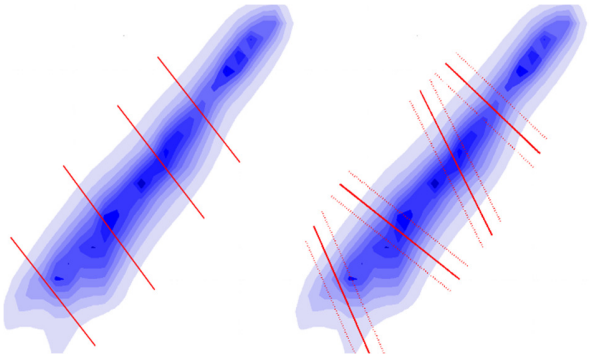


Figure 9. Density contours with temporal borders based on a linear tendency line (left) and with temporal borders and thresholds for smooth color gradients based on a curved tendency line (right).

Furthermore, thresholds for temporal borderlines can be applied for smooth color transitions. Different temporal surface segments carry different color hues. Within the same surface segment, the color hue remains the same but its intensity varies with the change of density.

In our approach, we adopted the "rainbow color scheme", which is essentially the visible and continuous electromagnetic spectrum. Its main color hues transit from red, orange, yellow, green, blue to violet. The spectrum can be divided into an arbitrary number of intervals. Users may easily anticipate and comprehend the color transitions. In our approach, we assign each time interval to a certain color hue – the medium color of the rainbow subinterval.

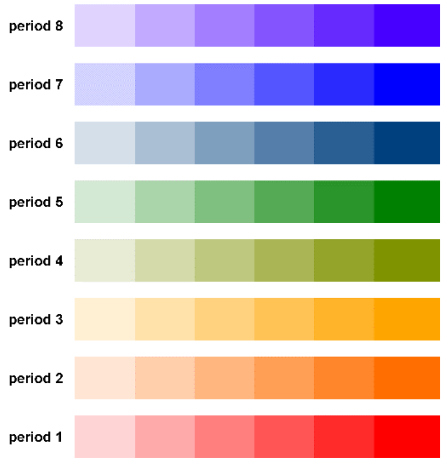


Figure 10. Rainbow color scheme.

Figure 10 exemplary illustrates eight different rainbow color hues with each being displayed in up to six different color intensities from light to dark. For instance, the red color scheme refers to time period 1 and contains six different red tones, which are related to six different density values/ value intervals. We split the entire time of our dynamic dataset into equal time intervals. The interval size can be determined based on the user's interest.

Hue represents time (e.g., discretized at 1 hour intervals) and color intensity corresponds to the density of observations (low intensity refers to low density and high intensity to high density). A continuous color scheme should not include more than three hues; otherwise the visual perception may suffer. An exception is the rainbow scheme. Most people know the differences in short and long wavelengths of visible light and are therefore familiar with the rainbow color gradation.

We decided to use the rainbow color scheme in order to fulfill the following two criteria:

'*Clear differentiation'*: colors of adjacent segments should be clearly distinguishable from each other. In particular, the brightness spectrum from low to high intensity should be distinct for each color hue from those of the others.

'*Continuity*': The color hues including their different intensities should represent the movement, thus, consists of a continuous color gradation. From the first hue allocated to the first temporal interval to the last one, the map user should be able to visually detect this continuity through a continuous color scheme. This color scheme has to be commonly known/familiar and intuitively understandable.

In literature, rainbow color maps are commonly used, but often are considered as harmful for continuous data [39]. The arguments against rainbow schemes include the inappropriateness for colorblind people, the appearing of divisions between hues, which lead to visual "edges" in the map, the meaningless spectral order of the hues and difficulties to recognize details. In particular to differentiate qualitative and quantitative attributes through polygon hue, the use of the rainbow scheme is often criticized. Figure 11

illustrates five different color scheme approaches for the STDmap using the rainbow color scheme (a) and four alternatives involving a color gradient from blue to purple (b) as well as three color scheme from Colorbrewer.org [40]: diverging (c), sequential multiple hues (d) and sequential single hue (e).

To visualize continuous data, often bipolar color illustration is used. On the other hand, also the rainbow color scheme is frequently used, for example to visualize the earth gravitational field (geoid anomalies) or illustrate weather-related intensities, such as storm severity [41]. Although the rainbow color scheme with its color gradation is commonly known, the continuity information based on color gradation might be easier to identify in the options provided in Figure 11-b,c,d,e. However, in Figure 11-c,d,e individual contour segments are difficult to identify due to the use of/color transition to white or yellow. Although the movement of the spatially extended object (SEO) in time is clearly visible in Figure 11-b, individual neighboring temporal segments could be confusing – which is not the case in Figure 11-a. To fulfill/combine the two contradictory criteria of the continuity and the clear differentiation, we have to make a compromise. The rainbow scheme might be more appropriate for some cases while a continuous color scheme involving 2-3 hues might suit better for other applications. For our STDmapping approach, we preferred to use the rainbow scheme. However, a user test is needed to verify the proper use of rainbow color scheme for our STDmap.

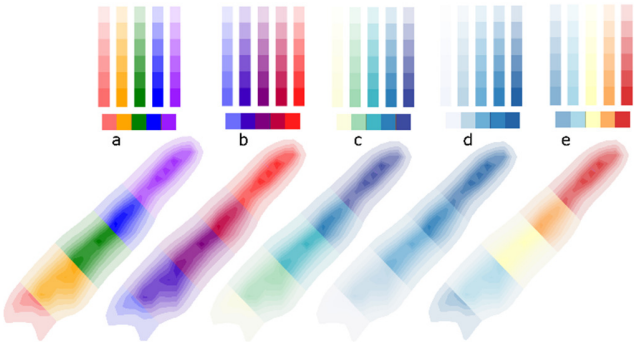


Figure 11. Different color scheme approaches.

With regard to the division of density surface by means of the temporal tendency lines (following either a straight or a curved route), we introduce the perpendicular lines to each tendency line as the temporal borders between the two neighboring time intervals of the underlying KDE map. The color transition between two temporal segments can be either abrupt or smooth. In case of smooth temporal borders, a defined threshold for the smooth color transition is set. The threshold refers to a certain time before and after the abrupt temporal borders. That leads to two parallel border lines – one to the left, the other to the right of the abrupt border line. The distance (time) between each smooth border line and the respective abrupt border line can be constant and variable.

Thus, our STDmapping approach provides a solution for the visual incorporation of temporal information within density surfaces of layered tints.

V. RESULTS AND DISCUSSION

For applying density visualization to our test dataset, containing lightning points during April 26th 2013, we used both our proposed STDmapping approach and the commonly used KDE method. Applying the latter traditionally for each spatio-temporal cluster, a segment of density map with layered tints was produced as illustrated in Figure 12, which however is not satisfying due to parts of overlays and occlusion. It leads to a loss of certain local and of the overall density information.

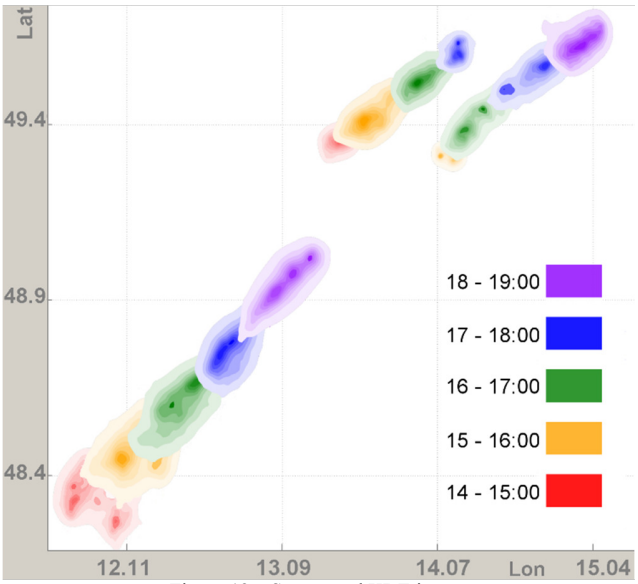


Figure 12. Segmented KDE in one map.

Applying transparency does not solve this drawback adequately. Due to the fact, that the density contour intervals of each temporal interval have the same hue but differ in color intensities, a transparency changes the intervals with low intensities to almost invisible – even if a contour border with a slightly more intense color is added. Hidden parts of overlapped contour intervals will still not be sufficiently recognizable.

Thus, using KDE maps for each spatio-temporal cluster provides only direct depiction of time for non-overlapping contour surfaces. Furthermore, visual exploration of density information in the overlapping parts is only possible for the surface on top; in case transparency is applied it is very difficult. Another disadvantage occurs when one is interested in density information including points detected shortly before and after the temporal interval border.

By applying the new STDmapping approach to our test dataset and following the workflow in Figure 6, we created eight different output maps (Figure 13 - Figure 20). The temporal borders were based on either straight (*A*) or curved

lightning cluster moving tendency lines (*B*). Furthermore, we used two different temporal thresholds: one hour, respectively 30 minutes. Moreover, we applied the abrupt and the smooth concept for color gradients between temporal segments.

A. STDmaps based on straight tendency lines

Four figures illustrate results for spatio-temporal density maps (STDmaps) based on straight tendency lines with the interval of one hour in Figure 13 and Figure 14 and 30 minutes in Figure 15 and Figure 16. The color gradients between temporal borders are abrupt in Figure 13 and Figure 15 and smooth in Figure 14 and Figure 16.

Using straight tendency lines, the temporal borders appear parallel to each other, in particular with abrupt color gradients. Larger distances between sequential temporal borders refer to a faster movement phase of the dynamic phenomenon, whereas the closer successive temporal borders indicate a slower movement of the lightning clusters.

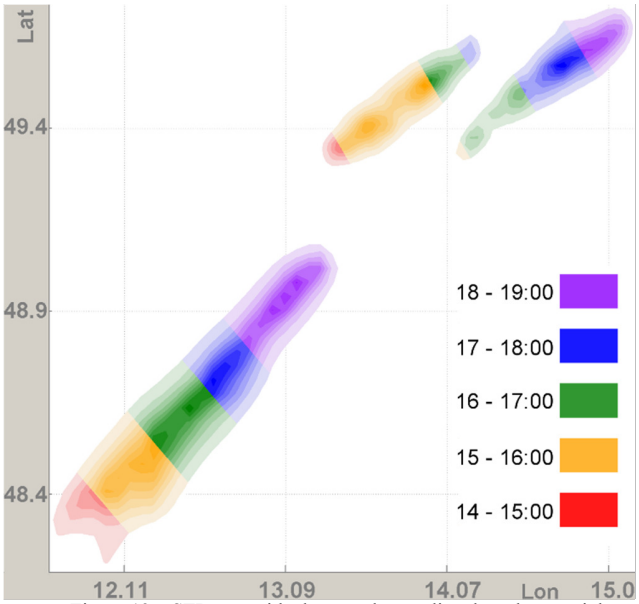


Figure 13. STDmap with abrupt color gradient based on straight tendency lines and the temporal interval of one hour.

When fewer temporal segments are used (e.g., five segments in Figure 13), the map reader may fast and easily extract the distinctive temporal information. When a larger number of temporal segments are used (e.g., in Figure 15), the map reader has to distinguish between more different colors referring to temporal information. This explorative interpretation becomes more effortful. On the other hand, more temporal segments reveal more details and may thus enable a more comprehensive insight in the dynamics of the data (e.g., temporal change of local point density).

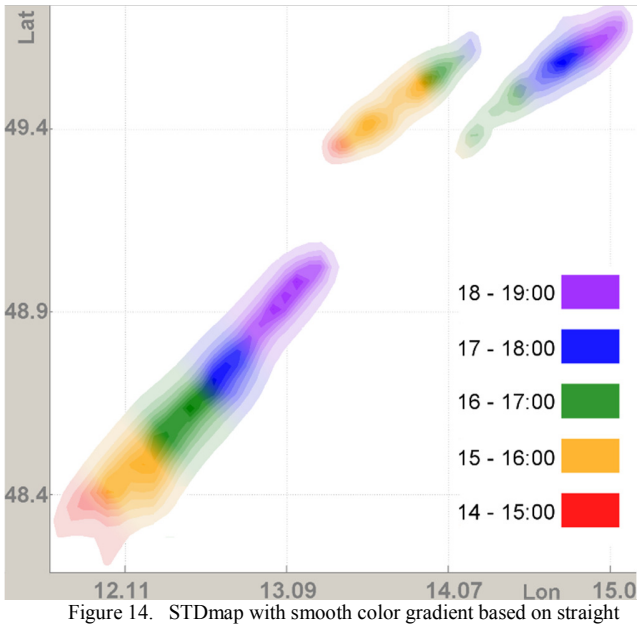


Figure 14. STDmap with smooth color gradient based on straight tendency lines and the temporal interval of one hour.

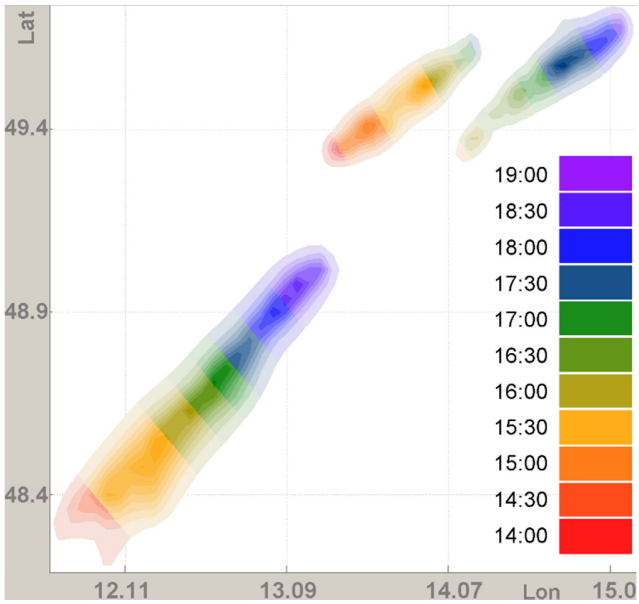


Figure 15. STDmap with abrupt color gradient based on straight tendency lines and the temporal interval of 30 minutes.

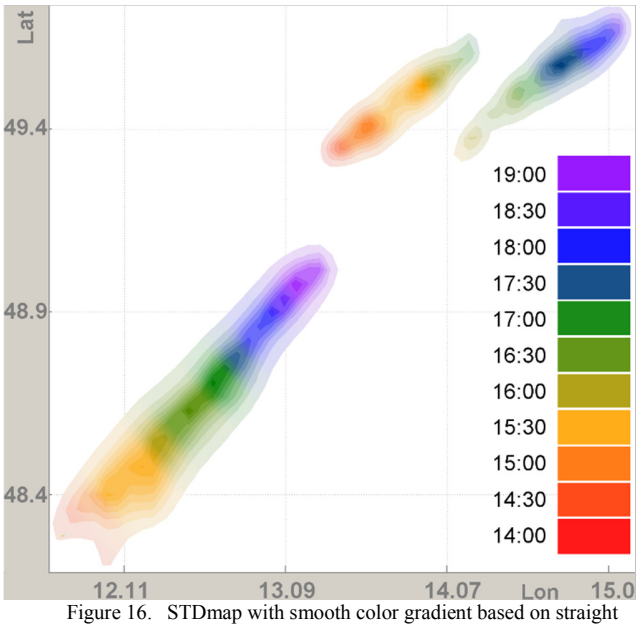


Figure 16. STDmap with smooth color gradient based on straight tendency lines and the temporal interval of 30 minutes.

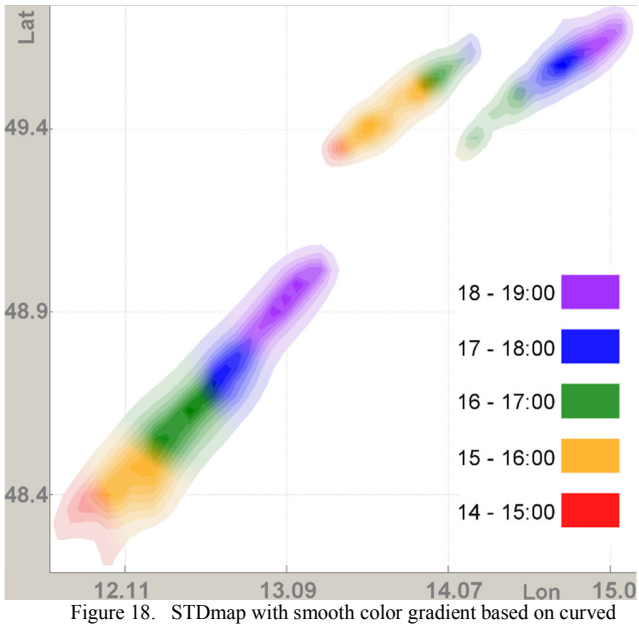


Figure 18. STDmap with smooth color gradient based on curved tendency line the temporal interval of one hour.

B. STDmap based on curved tendency lines

Four further figures illustrate the results for STDmaps based on curved tendency lines. The interval of one hour was used in Figure 17 and Figure 18 and 30 minutes for Figure 19 and Figure 20. The color gradients between temporal borders are abrupt in Figure 17 and Figure 19 and smooth in Figure 18 and Figure 20.

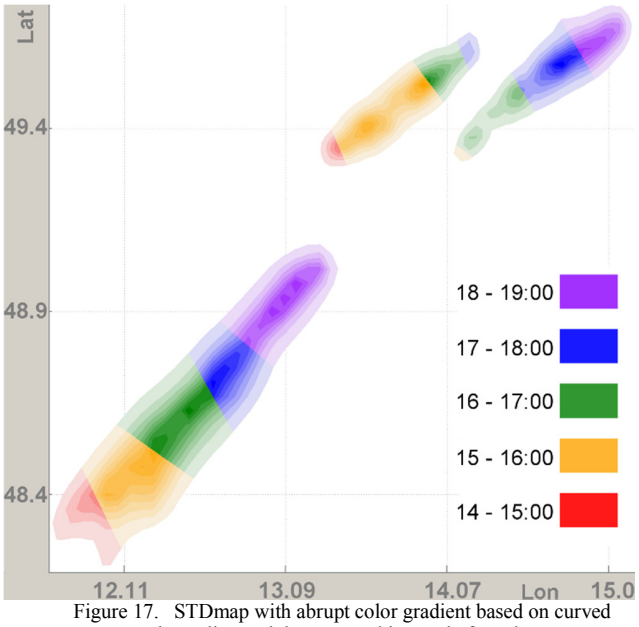


Figure 17. STDmap with abrupt color gradient based on curved tendency line and the temporal interval of one hour.

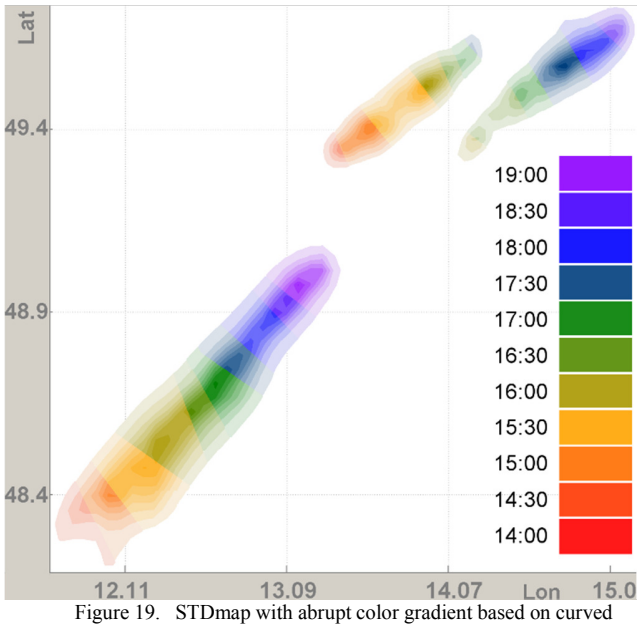


Figure 19. STDmap with abrupt color gradient based on curved tendency line and the temporal interval of 30 minutes.

A comparison of the results from the straight tendency line with those from the curved tendency line (e.g., Figure 14 with Figure 18 using five temporal segments or Figure 16 with Figure 20 using 11 temporal intervals) clearly shows the visual similarity of STDmaps, particularly in case of a rather large threshold for temporal borders.

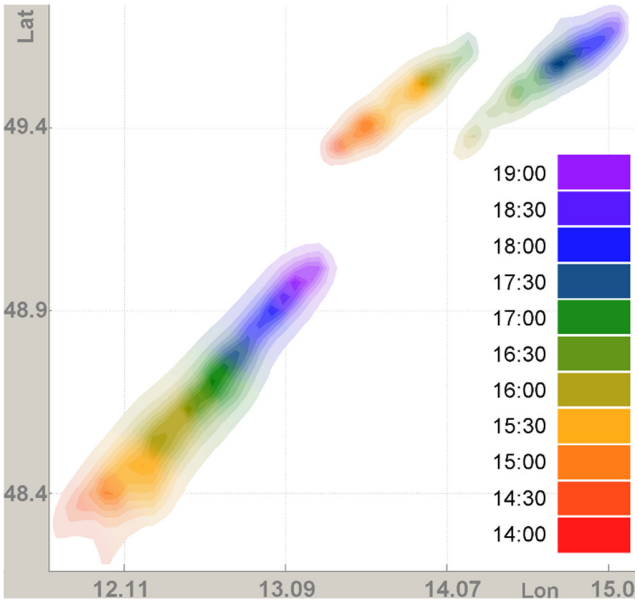


Figure 20. STDmap with smooth color gradient based on curved tendency line and the temporal interval of 30 minutes.

It can be obviously perceived in all STDmaps that the entire density information is clearly visible while the temporal information about phenomena dynamics in terms of speed and moving direction provides the added value. All lightning clusters are moving northeastwards. The upper left cluster is moving faster around 3:30pm than at any other time and it had two density peaks around 3pm and 4pm.

The clear-cut temporal cluster borders reveal another advantage: Density information in layered tints within each specific time interval is clearly visible and separable from neighboring segments. The smooth color transitions between neighboring segments are closer to the reality and correspond better to the visual perception: lightning points occurring for instance some minutes after 3pm can be located inside the 2-3pm segment and points appearing some minutes before 3pm might be placed inside the 3-4pm segment. With the help of an adaptive slider, the smoothing effect can be set for a small time interval (e.g., 14:55 – 15:05) or a large one (maximum smoothing interval: half of the time interval left and right of the temporal border, e.g., 14:30 – 15:30). Moreover, an elastic slider enables the use of different smoothing intervals adaptive to the cluster overlap and thus to the changing cluster speed. In our case we used a threshold of 10 minutes (five minutes before and after each abrupt border line). For an easy comprehension, we suggest to limit the number of colors (time intervals) to no more than about 15. In case of very extensive temporal range, the brightness of the same tone within the same interval can be adopted. For instance, 24 hours can be cut into six by four hours intervals. Within each interval four different brightness of the same tone can be used. In order to verify the proposed color mapping, an extensive user evaluation is necessary.

C. *Wrongly assigned points*

STDmapping approach is suitable for constantly moving SEOs. The approach creates a segmented contour interval for each track. However, the use of abrupt color gradients may lead to temporally wrongly assigned points. The tracking and in particular used clustering method (distance threshold) as well as the temporal segmentation (time intervals and tendency segmentation model) are decisive for the allocation of points to the temporal segments. These decisive steps (parameters) can be adapted to different moving situations along the trajectory. For example, in case a moving SEO changes its speed along the track, temporal intervals and smooth zones could be defined differently in order to reduce wrongly assigned points.

The following two figures illustrate the temporally wrongly assigned points (in black) detected in the STDmap for our test dataset while using a temporal interval of one hour and a linear tendency line. In Figure 21 abrupt temporal borders were used and in Figure 22 smooth zones (for smooth color transitions) of +/- 10 minutes were applied (semi-transparent pink polygons). The number of wrongly assigned points are displayed for each temporal border in Figure 21 and for each temporal interval between the smooth zones in Figure 22. We assume that the smooth zone visually refers to both adjacent temporal intervals. Out of altogether 6885 points for the ‘abrupt’ case 788 points (11.4 %) appeared to be assigned wrong and for the ‘smooth’ case 189 points (2.7 %) appeared to have the wrong temporal color code of the underlying STDmap. Thus, using the smooth STDmap fewer parts are visually allocated to the wrong real temporal interval.

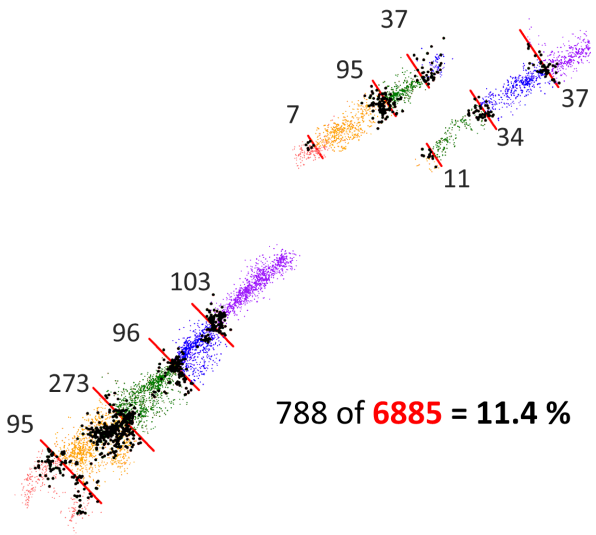


Figure 21. Wrongly assigned points (in black), using abrupt temporal borders.

A lower distance threshold leads to fewer dis-allocated points. The smaller the number of temporal intervals, the more cutlines are defined and probably more temporally false

allocated points occur. The wider the smooth area, the fewer the number of wrongly assigned points.

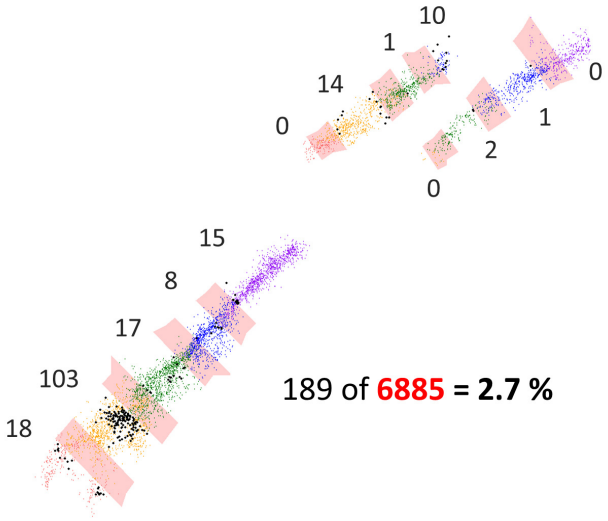


Figure 22. Wrongly assigned points (in black), using smooth temporal borders.

Basically, as the tendency line and segmented cuts are more generalized, the more wrongly assigned points occur. Thus, a spline tendency line should result in less wrongly assigned points than the use of a linear one. If a point is wrongly assigned, then it is mostly behind the segment border to the next temporal interval. A very slow movement consequently results in a higher number of wrongly assigned points. In case a SEO is moving for- and backwards or intersects a lot with itself, the produced contour intervals would overlap and thus map reading becomes more difficult, even if transparency is applied.

Between STDmapping and "KDE maps for each temporal interval" there are certain complementarities. We have shown that both approaches have their pros and cons. Our proposed STDmap is an alternative procedure to the overlapping KDE maps. Surely, there might be applications for which the use of KDE maps is more appropriate and others for which the STDmapping approach is more suitable. For our application case of moving lightning data, we prefer the use of STDmap. Comprehensive user tests are necessary to verify the right choice for different applications.

D. Limitations and comparison with existing approaches

Existing density mapping approaches are able to consider either two moments of time or a series of time intervals within density information visualization in a STC. However, the targeted visual communication of point density changes necessitates user interactions, especially when a cluster of interest is bounded by other clusters, it can be hardly explored without interaction. The 3D density STC suggested by Nakaya and Yano [25] is a comparable approach, where a series of time steps is taken into account within density information visualization with the aim to illustrate the change of point density in time. However, the density changes in time can only be explored interactively by panning, zooming and rotating etc. The STC approach does provide explicit spatio-temporal density information while the STDmapping method assigns some points to the wrong temporal intervals. However, the STC approach needs strong user interaction for the exploration of point density changes. If a cluster of interest is surrounded by other clusters, it can be hardly explored. Our approach has overcome this drawback by storing and presenting temporal information in different colors in a STDmap (in 2D). Spatial and temporal clustering parameters/thresholds can be adapted in order to improve the resulting STDmap. The approach creates appropriate segmented density surfaces. In other words, the approach is suitable if a movement and a main movement direction of the phenomena (polygon) are given and thus a tendency line together with temporal borderlines can be automatically identified. However, several movement cases may cause difficulties to identify the segments: (1) if the SEP is simultaneously expanding in several directions the tendency line needs to be split, which is no trivial task in the practice. (2) If the SEP is moving and returning after a circular track back to a previously passed location/area; or if the SEP is moving for and backwards. In these cases spatio-temporal polygons would overlap significantly and the resulting STDmap may become illegible.

E. Usability of STDmapping

To verify the usability of our approach a comprehensive test of how users interact with the STDmapping is necessary. It deals with a multidisciplinary research field and requires knowledge about the user, the user's task and the involved technology [42]. The target users of our approach are domain specialists who interact with the visualized dynamic phenomena and who need to identify spatio-temporal changes of local and global point densities. The anticipated user tests aim to investigate how the visualized information is perceived and understood. The usability of alternative STDmaps can be compared or iteratively improved. The iteration bears a two-fold meaning. On the one hand, the STDmap designers benefit from user's behavior. On the other hand, an improved STDmap will better empower the users. In the latter case, the users get trained to get along with the STDmapping approach. Since the interactive and explorative use of the STDmaps of dynamic SEPs requires some vocational adjustment, the corresponding usability tests should be conceptualized as a long-term endeavor involving repeating test sessions with the same target users.

VI. Conclusion and Future Work

Visualizing density and distribution information is a key support for the understanding of spatio-temporal phenomena represented by point data. However, the temporal information is not yet adequately handled in existing density mapping approaches. Our work has closed this research gap by incorporating and visualizing the temporal change of point cluster in a 2D density map. Our approach is termed as STDmapping according to which a density surface of layered tints can be divided into different temporal segments. Each

segment is then visualized by a color hue with varying intensities. The temporal borders are visualized as lines perpendicular to the moving tendency lines. Tendency lines are either straight or curved. Moreover, abrupt or smooth color gradients between neighboring temporal segments can be applied.

The resulted STDmaps comprise spatio-temporal information about density, distribution, movement patterns such as moving direction and speed of dynamic point clusters. Thus, our approach supports the pattern detection/extraction of spatio-temporal phenomena without having to activate interactive tools. Furthermore, our approach can be adapted to dynamic phenomena represented by other point events as well as to moving point groups (e.g., animal swarms).

In the next step of our work, the usability of STDmap for lightning data will be investigated. It requires the participation of users who are domain specialists and should make decisions based on their understanding of visualized lightning data. Specific user tasks related to the extraction of certain spatio-temporal density information or dynamic patterns should be repeatedly conducted and evaluated. Various interactive functions should be made available to allow these users to manipulate the visualization for the purpose of a more efficient exploration, for instance, by adapting the color scheme, changing the time interval, etc.

Meanwhile, we plan to investigate the relation between the characteristics of initial data (density, distribution, spatio-temporal change of point coordinates) and their modeling parameters (movement tendency, time interval, boundary lines) with the purpose to describe the dynamic phenomena with minimum information loss or distortion for the subsequent visualization and use of STDmaps. Furthermore, an adaption of our approach for 3D point data is also possible. Last but not least, it is worthwhile to develop the dynamic mapping technologies for geo-sensory systems, which, for example, demand the dynamic derivation of density layers, contour lines or discrete classes from the values regularly sent by various sensors.

ACKNOWLEDGMENT

The authors gratefully acknowledge Nowcast Company for providing lightning test dataset and the support of the Graduate Center Civil Geo and Environmental Engineering at Technische Universität München, Germany.

REFERENCES

[1] S. Peters and L. Meng, "Spatio Temporal Density Mapping of a Dynamic Phenomenon," in *GEOprocessing 2014 - The Sixth International Conference on Advanced Geographic Information Systems, Applications, and Services*, Barcelona, Spain, 2014, pp. 83-88.

[2] N. Andrienko and G. Andrienko, "Visual analytics of movement: An overview of methods, tools and procedures," *Information Visualization,* vol. 12, pp. 3-24, 2013.

[3] W. A. Mackaness, A. Ruas, and L. T. Sarjakoski, *Generalisation of geographic information: cartographic modelling and applications*. Amsterdam, The Netherlands: Elsevier Science, 2007.

[4] K. Romanenko, D. Xiao, and B. J. Balcom, "Velocity field measurements in sedimentary rock cores by magnetization prepared 3D SPRITE," *Journal of Magnetic Resonance,* pp. 120-128, 2012.

[5] S. Stoilova-McPhie, B. O. Villoutreix, K. Mertens, G. Kemball-Cook, and A. Holzenburg, "3-Dimensional structure of membrane-bound coagulation factor VIII: modeling of the factor VIII heterodimer within a 3-dimensional density map derived by electron crystallography," *Blood,* vol. 99, pp. 1215-1223, 2002.

[6] J. W. Tukey, *Exploratory data analysis*: Addison-Wesley, 1977.

[7] B. W. Silverman, *Density estimation for statistics and data analysis* vol. 26: CRC press, 1986.

[8] N. Cressie, "Statistics for spatial data," *Terra Nova,* vol. 4, pp. 613-617, 1992.

[9] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization* vol. 383: Wiley. com, 2009.

[10] D. O'Sullivan and D. J. Unwin, *Geographic information analysis*: John Wiley & Sons, 2003.

[11] M.-P. Kwan, "Geovisualisation of activity-travel patterns using 3D geographical information systems," in *10th international conference on travel behaviour research (pp. pages pending), Lucerne*, 2003, pp. 185-203.

[12] I. Assent, R. Krieger, E. Müller, and T. Seidl, "VISA: visual subspace clustering analysis," *ACM SIGKDD Explorations Newsletter,* vol. 9, pp. 5-12, 2007.

[13] J. M. Krisp and O. Špatenková, "Kernel density estimations for visual analysis of emergency response data," in *Geographic Information and Cartography for Risk and Crisis Management*, ed: Springer, 2010, pp. 395-408.

[14] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani*, et al.*, "A visual analytics approach to understanding spatiotemporal hotspots," *Visualization and Computer Graphics, IEEE Transactions on,* vol. 16, pp. 205-220, 2010.

[15] C. F. Schmid and E. H. MacCannell, "Basic Problems, Techniques, and Theory of Isopleth Mapping*," *Journal of the American Statistical Association,* vol. 50, pp. 220-239, 1955.

[16] T. A. Slocum, R. B. McMaster, F. C. Kessler, and H. H. Howard, *Thematic cartography and geovisualization*: Pearson Prentice Hall Upper Saddle River, NJ, 2009.

[17] M. Langford and D. J. Unwin, "Generating and mapping population density surfaces within a geographical information system," *The Cartographic Journal,* vol. 31, pp. 21-26, 1994.

[18] E. M. Jansenberger and P. Staufer-Steinnocher, "Dual Kernel density estimation as a method for describing spatio-temporal changes in the upper austrian food retailing market," in *7th AGILE Conference on Geographic Information Science*, 2004, pp. 551-558.

[19] J. M. Krisp and S. Peters, "Directed kernel density estimation (DKDE) for time series visualization," *Annals of GIS,* vol. 17, pp. 155-162, 2011 2011.

[20] J. M. Krisp, S. Peters, and F. Burkert, "Visualizing Crowd Movement Patterns Using a Directed Kernel Density Estimation," in *Earth Observation of Global Changes (EOGC)*, Munich, Germany, 2013, pp. 255-268.

[21] J. M. Krisp, S. Peters, C. E. Murphy, and H. Fan, "Visual Bandwidth Selection for Kernel Density Maps,"

Photogrammetrie - Fernerkundung - Geoinformation, vol. 2009, pp. 445-454, 2009/11/01/ 2009.

[22] J. M. Krisp, S. Peters, and M. Mustafa, "Application of an Adaptive and Directed Kernel Density Estimation (AD-KDE) for the Visual Analysis of Traffic Data," in *GeoViz2011*, Hamburg, Germany, 2011.

[23] J. M. Krisp and S. Peters, "Visualizing Dynamic 3D Densities: A Lava-lamp approach," in *13th AGILE International Conference on Geographic Information Science*, Guimaraes, Portugal, 2010, pp. 10-14.

[24] S. Peters and J. M. Krisp, "Density calculation for moving points," in *13th AGILE International Conference on Geographic Information Science*, Guimaraes, Portugal, 2010, pp. 10-14.

[25] T. Nakaya and K. Yano, "Visualising Crime Clusters in a Space time Cube: An Exploratory Data analysis Approach Using Space time Kernel Density Estimation and Scan Statistics," *Transactions in GIS,* vol. 14, pp. 223-239, 2010.

[26] C. F. Schmid and E. H. MacCannell, "Basic Problems, Techniques, and Theory of Isopleth Mapping*," *Journal of the American Statistical Association,* vol. 50, pp. 220-239, 1955.

[27] U. Demšar and K. Virrantaus, "Space–time density of trajectories: exploring spatio-temporal patterns in movement data," *International Journal of Geographical Information Science,* vol. 24, pp. 1527-1542, 2010.

[28] R. Scheepens, N. Willems, H. van de Wetering, G. Andrienko, N. Andrienko, and J. J. van Wijk, "Composite density maps for multivariate trajectories," *Visualization and Computer Graphics, IEEE Transactions on,* vol. 17, pp. 2518-2527, 2011.

[29] N. Willems, H. Van De Wetering, and J. J. Van Wijk, "Visualization of vessel movements," in *Computer Graphics Forum*, 2009, pp. 959-966.

[30] N. Willems, H. Van De Wetering, and J. J. Van Wijk, "Visualization of vessel movements," in *Computer Graphics Forum*, 2009, pp. 959-966.

[31] G. McArdle, A. Tahir, and M. Bertolotto, "Interpreting map usage patterns using geovisual analytics and spatio-temporal clustering," *International Journal of Digital Earth,* pp. 1-24, 2014.

[32] P. Forer and O. Huisman, "Space, time and sequencing: Substitution at the physical/virtual interface," *Information, Place, and Cyberspace: Issues in Accessibility,* pp. 73-90, 2000.

[33] G. Andrienko and N. Andrienko, "A general framework for using aggregation in visual exploration of movement data," *Cartographic Journal,* vol. 47, pp. 22-40, 2002.

[34] H. D. Betz, K. Schmidt, and W. P. Oettinger, "LINET – An International VLF/LF Lightning Detection Network in Europe," in *Lightning: Principles, Instruments and Applications*, H. D. Betz, U. Schumann, and P. Laroche, Eds., ed Dordrecht: Springer Netherlands, 2009, pp. 115-140.

[35] S. Peters, H.-D. Betz, and L. Meng, "Visual Analysis of Lightning Data Using Space–Time-Cube," in *Cartography from Pole to Pole - 26th International Cartographic Conference (ICC)*, Dresden, Germany, 2013, pp. 165-176.

[36] S. Peters and L. Meng, "Visual Analysis for Nowcasting of Multidimensional Lightning Data," *ISPRS International Journal of Geo-Information,* vol. 2, pp. 817-836, 2013.

[37] S. Peters, L. Meng, and H. D. Betz, "Analytics approach for Lightning data analysis and cell nowcasting," in *EGU General Assembly Conference Abstracts*, 2013, pp. 32-3.

[38] C. De Boor, *A practical guide to splines* vol. 27: Springer-Verlag New York, 1978.

[39] D. Borland and R. M. Taylor II, "Rainbow color map (still) considered harmful," *IEEE Computer Graphics and Applications,* vol. 27, pp. 14-17, 2007.

[40] C. Brewer and M. Harrower. (2014, 10/2014). *colorbrewer*. Available: http://www.colorbrewer2.org/

[41] R. Wicklin. (2013, 10/2014). *How to choose colors for maps and heat maps*. Available: http://blogs.sas.com/content/iml/2014/10/01/colors-for-heat-maps/

[42] M. Haklay, *Interacting with geospatial technologies*: Wiley Online Library, 2010.

Introducing a General Multi-Purpose Pattern Framework: Towards a Universal Pattern Approach

Alexander G. Mirnig and Manfred Tscheligi
Center for Human-Computer Interaction
Christian Doppler Laboratory for "Contextual Interfaces"
Department of Computer Sciences, University of Salzburg
Salzburg, Austria
Email: {firstname.lastname}@sbg.ac.at

Abstract— **Patterns have been successfully employed for capturing knowledge about proven solutions to reoccurring problems in several domains. Despite that, there is still little literature regarding pattern generation or common pattern quality standards across the various domains available. This paper is an extended version of a short paper presented at PATTERNS 14 [1], in which we introduced an attempt for a universal (i.e., domain independent) pattern framework. Via basic set theory, it is possible to describe pattern sets that are composed of several subsets regarding pattern types, quantities, sequence, and other relevant factors. This further enables us to describe patterns as sets of interrelated elements instead of isolated entities, thus corresponding with the scientific reality of complex problems with multiple relevant factors. The framework can be used to describe existing pattern languages and serve as a basis for new ones, regardless of the domain they are or were created for.**

Keywords-patterns; basics on patterns; pattern framework; set theory; pattern modeling

I. INTRODUCTION

Patterns have been used as a tool for capturing knowledge about proven solutions to reoccurring problems in a multitude of domains and disciplines. Most prominent among these are architecture, design, and software engineering [1][2][3][4][5]. Patterns allow documenting knowledge about methods and practices in a structured and systematic manner and can, therefore, serve as a valuable knowledge transfer tool within or even across disciplines. Another related benefit of patterns is that they can serve to "make implicit knowledge explicit" [6], i.e., they can be used to explicitly capture what is normally only acquired via experience after having worked in a certain field or domain for an extended period of time. They can thus go beyond and supplement the "raw" information contained in guidelines with a more solution- and practice-oriented dimension. The information contained in such patterns can then be provided to others (researchers or other interested parties) in a relatively quick and efficient manner, as it contains information about solutions that are already proven to work.

Having access to a structured collection of information from implicit and explicit knowledge about research practices is useful for any domain in which research is conducted. So it would make sense to extend the pattern approach or even establish patterns as a general field of basic research, with extensions into particular domains and disciplines. This wider potential of patterns has been recognized and has been summarized by Borchers [7] in the following way: "There is no reason why experience, methods or values of any application domain cannot be described in pattern form as long as activity includes some form of design, creative or problem-solving work." Despite this, there is little general (i.e., domain independent) literature available on patterns and pattern finding or creation. This is not a new idea [8] and there have been efforts to go deeper into the commonalities of various pattern approaches and patterns in general by, e.g., the work of Meszaros and Doble [9] and Winn and Calder [10].

Two of the main benefits of patterns are that they facilitate re-application of proven solutions and that they serve to make implicit knowledge explicit. These benefits are of particular importance to those, who do not already have this knowledge themselves, i.e., it is a way to draw from a vast pool of knowledge that would otherwise be gained via experience, over a long period of time. If working with patterns has extensive domain experience as a prerequisite, then those that would need that knowledge the most would arguably benefit the least from it.

We argue for a general strand of research on patterns as a means to capture knowledge about research practices. With such a theoretical basis available, practitioners from any domain could have a pool of knowledge to draw from, which would help them create patterns suitable for their needs. This should not mean that a variety in pattern languages and approaches is not desirable. It makes sense to assume that different domain requirements need different pattern approaches. However, the basics of patterns should ideally be similar for everyone and easily accessible, like, e.g., with general mathematics. A statistician needs and employs different mathematical means than a fruit vendor. But both draw from the same pool of general mathematics as their basis. In our research, we try to look at patterns in a similar way. We want to promote their use as a universal tool to structure knowledge in all kinds of areas and disciplines.

In this paper, we will take a look at pattern approaches in general and the commonalities between them. We will then integrate these into a formal pattern framework, with the aim

of providing a formally sound and flexible basis, which allows practitioners and researchers to create their own patterns and pattern collections within their respective domains. To this end, we pursue four main goals in developing our framework:

- the framework should be a suitable basis for and, therefore, be compatible with most (if not all) existing pattern approaches and languages
- it should contain basic functionalities that allow meaningful structuring and referencing of patterns
- the framework must dictate the pattern content only in the most rudimentary way, so that it is not restricted to only one or very few disciplines
- the framework must be formally sound but also easy to work with, so that it can be applied by large number of individuals

The final goal of this research is to arrive at a structured but still easy to understand framework that captures the essence of patterns and makes them understandable as well as usable for practitioners and researchers in any domain. We do this via a basic set theoretic [11] analysis that allows describing patterns and pattern languages in a general manner. Such a general analysis of patterns allows us to treat them as separate phenomena, independent of the domains they are created and used in. Set theory is one of the most basic, but at the same time very powerful, mathematical tools available. By using set theory, we can ensure consistency of our framework, while still keeping things basic and relatively easy to understand. An additional benefit of our approach is that it permits the creation of pattern sets across different pattern languages that address a similar purpose. This can facilitate the consolidation of already existing knowledge within the various domains.

This set theoretic framework serves as a domain independent basis for reflections on how patterns can or should be created and structured. It can be extended to fit the needs of a particular discipline or area, if that would become necessary, but is, at its core, a purely formal tool that is not restricted to any domains or disciplines. In this paper, we begin with an overview of existing general literature on patterns in Section II, followed by some explanations regarding the basics of set theory and why we deem it a suitable tool for the purpose of this paper. In Sections V and VI, we provide an outline of the proposed set theoretic pattern framework. In Section VII, we supplement the framework with general recommendations on how to find patterns for multidisciplinary applications of the framework. In Section VIII, we present an example application of the framework to structure an existing pattern collection. In Section IX, we discuss limitations and future work potentials of the framework, with a brief conclusion at the end in Section X.

II. Related Work

Patterns have been employed in a multitude of application domains [1][4][12] and a good number of extensive pattern collections [7][13][14] have been created in the past. Literature on the pattern generation process itself, sometimes also referred as *pattern mining* [13], is still scarce [15]. Existing literature on pattern generation is mostly focused on specific domains [7][4][8][12]. The work of Gamma et al. [13] can be considered important elementary literature, but it is still centered on software design. Although covering a wide spectrum of software design problems, it is arguably of limited applicability outside of the software engineering domain. The same can be said for other specialized pattern generation guidance [8], which would require adaptation to be employed in other domains (e.g., biology or linguistics). Falkenthal et al. [16] introduced a promising approach for validating solution implementations of patterns in various domains, though provided only one nontraditional use case (Costumes in Films) for their approach.

The advantages of patterns would be both desirable and feasible [7] for these other domains. Vlissides [17] provides a good summary of what patters can and cannot do. The perceived advantages of patterns might be summed up as follows:

- they capture expertise and make it accessible to non-experts
- their names collectively form a vocabulary that helps developers communicate better.
- they help people understand a system more quickly when it is documented with the patterns it uses.
- they facilitate restructuring a system whether or not it was designed with patterns in mind.

Another interesting aspect of patterns is that one single pattern is usually not enough to deal with a certain issue. Alexander [2] himself already expressed this by stating the possibility of making buildings by “stringing together patterns.“ However, the pattern itself does not always include the information of which other pattern might be relevant in a particular case. This information is only available once the pattern is part of an actual pattern language of several related patterns. Borchers [7] introduced the notion of high level patterns, which reference lower level patterns to describe solutions to large scale design issues. This hierarchy is expressed via references in the patterns themselves. Borchers’ view of high and low level patterns is a good way of understanding and describing patterns as interconnected entities. A suitable framework for patterns and pattern languages should ideally be able to capture the – sometimes complex – relations between patterns and allow mapping of individual patterns to higher level or overarching problems or goals.

One concept that is similar to the ideas pursued in this paper is that of ontologies. While term ‘ontology’ itself can have several meanings, the following short definition by Blomqvist and Sandkuhl [18] provides a good summary of how the term is usually understood in ontology engineering: “An ontology is a hierarchically structured set of concepts describing a specific domain of knowledge that can be used to create a knowledge base. An ontology contains concepts, a subsumption hierarchy, arbitrary relations between concepts, and axioms. It may also contain other constraints and functions.” Given this description, one might assume that ontologies would be an ideal tool to capture and transfer

domain knowledge universally. However, there are two limitations that make ontology engineering approaches run counter to the goals pursued in this paper. While ontologies can promote the application of good practices [20][21][22], reusability of ontologies is still considered a serious, and as of yet unsolved, challenge in ontology engineering [20][23][24]. Second, actually building an ontology is a very difficult tasks with many potential pitfalls, even for experts [20][21].

Both of these are serious limitations when considering suitability for users from a wide range of disciplines and/or skill levels as well as reusability of existing solutions. Patterns, thanks to their focus on reusability and the problems themselves instead of the abstract structure of a domain or field, seem more well-suited for the problems at hand.

Efforts to provide a general basis for patterns include the work of Meszaros and Doble [9], who developed a pattern language for pattern writing, which serves to capture techniques and approaches that have been observed to be particularly effective at addressing certain reoccurring problems. Their *patterns for patterns* were divided into the following five sections: Context-Setting Patterns, Pattern Structuring Patterns, Pattern Naming and Referencing Patterns, Patterns for making Patterns Understandable, Pattern Language Structuring Patterns. Another interesting approach being quite similar in its aims to the one presented in this paper is the *Pattern Language for Pattern Language Structure* by Winn and Calder [10]. They identified a common trait among pattern languages (i.e., they are symmetry breaking) and built a rough, nonformal general framework for pattern languages in multiple domains. These ideas are similar in concept to what we pursue in our research. The difference is that we want to provide a purely formal framework without or minimal statements regarding its content (such as types or traits). We want to focus on the basics behind patterns and structure these so that they can be applied as widely as possible, although we draw from their work (and that of others) to supplement the framework with general recommendations for pattern finding later in Section VII.

III. Set Theory – a Brief Introduction

Patterns, despite the term having a rather well established meaning, come in many shapes and forms and can be of varying complexity and verbosity. At the most basic level, they still have one thing in common – They consist of a number of statements, which are divided into several different categories of statements (e.g., pattern name, scenario, problem statement). A pattern is naturally much more than that, but this rather simple and elementary commonality is sufficient to begin building a framework from. A framework is a structure, an empty container that facilitates working with its contents (whatever these might be) in a consistent and organized way. In our case, this container should facilitate organizing and referencing patterns. Set theory is a mathematical method that allows organization of objects or data into so-called sets. Thus, if we understand patterns as collections statements in different categories, the connection to set theory becomes evident when we replace the word 'categories' with 'sets'. In the following paragraphs, we will outline the basics of set theory and highlight some of its advantageous attributes that we will use to build the pattern framework.

A set is an abstract, mathematical entity that contains other entities. These contents can either be sets themselves or singular, irreducible entities – so-called *elements*. Sets that are themselves contained in another set are called *subsets* of the set(s) they are part of (their *supersets*). Let us illustrate these considerations via the following example set S:

$$S = \{a, \{b, c\}\} \tag{1}$$

Sets are commonly denoted via curly brackets ('{' and '}') . In (1), we see two such sets: The set $\{b, c\}$ and the set $\{a, \{b, c\}$. The former is contained in and thus a subset of the latter. Therefore, we can say that $\{b, c\}$ is a subset of S and that S is a superset of $\{b, c\}$. This can more briefly be expressed via the symbols '$\subseteq$' and '$\supseteq$' in the following way: $\{b, c\} \subseteq S$; $S \supseteq \{b, c\}$. A subset is a *proper* subset if it is contained in another set, but there is at least one element that is part of the superset, but not the subset. In our example, a would be such an element, which is why we can furthermore state that $\{b, c\}$ is a proper subset of S. We write this as: $\{b, c\} \subset S$. The fact that a is an element of S can be expressed via the symbol '$\in$' in a similar fashion as: $a \in S$. A set of non-empty subsets is called a *partition*, if each element of the superset lies in exactly one element of the set of subsets. The set $\{\{a\}, \{b, c\}, \{d\}\}$ is a valid partition of the set $\{a, b, c, d\}$.

Sets are defined by their contents and one set is identical to another if every element of the former is also an element of the latter. The order of these elements does not matter. Consider the following examples:

$$\{b, c\} = \{c, b\} \tag{2}$$

$$\{a, \{b, c\}\} \neq \{b, c\} \tag{3}$$

The sets $\{b, c\}$ and $\{c, b\}$ both contain the same elements, so they are identical. The variable a is contained in $\{a, \{b, c\}\}$ but not in $\{b, c\}$, so these two sets are not identical. Via rather simple operations we can distinguish sets from each other via their contents and make statements regarding these same contents. These operations can be used to distinguish patterns from each other and structure their contents via subsets.

Two other very useful aspects of set theory, which we will employ later on, are ordered sets and the empty set. As mentioned before, the order of the elements in a set is usually irrelevant. Ordered sets can be used to arrange contents in a certain order. To distinguish them from regular sets, they are denoted by angle brackets ('<' and '>'). On a technical level, ordered sets are regular sets, where the order is determined by the number of subsets a certain element is part of. So the

ordered set <*a*, *b*, *c*> would really be the regular set {{*a*}, {*a*, *b*}, {*a*, *b*, *c*}}. *a* is contained in three subsets, *b* in two, and *c* only in one, which results in the order of *a*, *b*, *c*. Let us illustrate this further via the following examples:

$$<a, b> \neq <b, a> \quad (4)$$

$$\{\{a\}, \{a, b\}\} \neq \{\{b\}, \{b, a\}\} \quad (5)$$

$$\{\{c, a\}, \{b, a, c\}, \{c\}\} = <c, a, b> \quad (6)$$

The formula in (4) demonstrates that order is essential in ordered sets and (5) shows why that is the case, since the regular sets in (5) are simply the equivalents to the ordered sets in (4). (6) is meant to emphasize, that the order of the elements in the regular sets does not matter, only their frequency of occurrence does. The utility of being able to put things into sequence is very useful and we can use this to capture problems and the sequence they occur in. It can also be used to map the hierarchy of patterns as sequences from higher to lower level patterns.

A set is an abstract entity. It is defined by its contents but not identical to its contents. It is more than the sum of its elements. Therefore, the following is true:

$$a \neq \{a\} \quad (7)$$

While *a* is an element of the set that contains only *a*, it is *not* identical to that set. The set itself is a separate entity. An interesting consequence of this is that we can talk about sets, regardless of whether they actually contain anything. There is exactly one set, which does not contain any elements, and it is called the empty set (usually denoted by '∅' or '{}'). The empty set is a very versatile tool and can be used to, e.g., handle blank fields in a pattern (e.g., an incomplete pattern without keywords)

At this point, we should mention again that the pattern framework is intended for a wide range of individuals from all kinds of backgrounds. Depending on the reader's background, this section might have seemed either a bit complicated, or rather elementary for what can be considered a substantial section of this paper. The most important things to keep in mind for now are:

- regular sets to cluster and organize information
- ordered sets to put information into sequence
- the empty set to handle empty categories

And with that, we have covered elementary set theory to a sufficient degree, so that we can now shift our focus onto the patterns themselves.

IV. Patterns in General

A. Patterns and Pattern Sets

Now that we have provided an overview of elementary set theory and the components that we intend to employ, we want to go more into detail regarding patterns, pattern languages, as well as some of the concepts behind them. Beginning with the basic term "pattern", Alexander [2] characterized patterns in that "each pattern describes a problem that occurs over and over again in our environment, and then describes the core of the solution to that problem, in such a way that you can use this solution a million times over, without ever doing it the same way twice." In order to fulfill this requirement, patterns are usually held to some minimum standards of what they and their structure should contain. Probably the most common one is the structure suggested by Borchers [7] and van Velie [5], who suggest six main pattern elements: name, forces, problem, context, solution, and examples. Existing pattern structures come in a variety of levels of complexity – from detailed ones with a large number of subcategories [4] to comparably simplistic ones [12] with a small number of subcategories. So on a basic level, patterns can be understood as a structured assortment of statements. The statements contained in each instance of such a structure form a whole that provides a solution to a specific problem, describing both in detail to facilitate reapplication of the solution. It is this structure that we build into a framework via set theory.

A pattern language is a complete hierarchy of patterns, ordered by their scope [12]. Patterns can be divided into high- and low-level patterns [7], depending on the scale of the problem and its solution. High level patterns are more abstract and deal with larger scale problems. Low level patterns are more concrete and focus on smaller problems or parts of problems. In that sense, low level problems can be part of high level problems, which means that low level patterns can address or further specify problems from high level patterns. The distinction between high and low level patterns is not a strict one and depends on the respective individual's or group's perception of the degree of abstraction of a certain problem. In software engineering, the lowest level patterns are also referred to as *idioms*. Idioms contain very concrete solutions – mostly actual code snippets – that address very small scale coding problems.

In addition to "regular" patterns, there are also other types of patterns. The most notable of these are anti-patterns, which differ from regular patterns in that they do not describe a proven working solution to a problem, but rather a solution that is proven *not* to work or not work well. They follow the same structures as regular patterns, but instead of best practices they describe bad practices that should not be replicated. Since they describe solutions to reoccurring problems just as regular patterns do (with the delineating factor that the described solution should be avoided instead of replicated), a framework suitable for regular patterns should also be suitable for anti-patterns. In the framework, it will be possible to represent all these different pattern types.

V. The Basic Framework

We now translate these concepts into a basic set theoretic structure. We do so by employing mostly regular sets, ordered sets, and subsets. Note that the following analysis would work for any number of statements that is or can be structured in a similar way, which includes most, if not all, pattern structure (such as, e.g., the design patterns template laid out by Gamma et al. [4]). We chose to base our analysis on a concrete example, a Contextual User Experience (CUX)

pattern structure by Krischkowsky et al. [8] (see Table 1), in order to give a more current and less software engineering centered example. As noted previously, the following set theoretic analysis would apply to any other similarly structured assortment of statements (in this case most – if not all – imaginable pattern structures).

TABLE I. CUX PATTERN STRUCTURE [8]

Instructions on Each Pattern Section		
#	***Section Name***	***Instruction on Each Section***
1	Name	*The name of the pattern should shortly describe the suggestions for design by the pattern (2-3 words would be best).*
2	UX Factor	*List the UX factor(s) addressed within your chosen key finding (potential UX factors listed in this section can be e.g., workload, trust, fun/enjoyment, stress...). Please underpin your chosen UX factor(s) with a definition.*
3	Key Finding	*As short as possible - the best would be to describe your key finding (either from an empirical study or findings that are reported in literature) in one sentence.*
4	Forces	*Should be a detailed description and further explanation of the result.*
5	Context	*Describe the detailed context in which your chosen key finding is extracted/gathered from.*
6	Sugges-tions for Design	*1) Can range from rather general suggestions to very concrete suggestions for a specific application area. 2) The design suggestions should be based on existing knowledge (e.g., state of the art solutions, empirical studies, guidelines, ...). 3) More than one suggestion are no problem but even better than only one. 4) There can also be a very general suggestions and more specific "sub-suggestions".*
7	Example	*Concrete examples underpinned by pictures, standard values etc. Examples should not provide suggestions (this is done in the suggestion part) but rather underpin and visualize the suggestion presented above.*
8	Key-words	*Describe main topics addressed by the pattern in order to enable structured search.*
9	Sources	*Origin of the pattern (e.g., literature, other pattern, studies or results)*

The pattern structure shown in Table I consists of nine categories. This means that each pattern generated in this structure will consist of various statements in each of these nine categories. We now want to generate an actual pattern language set, let us call it CUX Language (and refer to it as *CL* for brevity's sake), based on the structure outlined in Table 1. We can do so by introducing nine subsets (i.e., sets of the set *CL*) CL_1 to CL_9, each subset corresponding to one of the nine categories (from Name to Sources, respectively) described above. This is what the initial structure looks like:

$$CL: \{CL_1, CL_2, \ldots, CL_9\} \quad (8)$$

At this very basic level, a pattern structure is nothing more than a set consisting of a number of (as of yet empty) subsets. To distinguish the *CL*-subsets from each other, they can be filled with a sequence of numbers reflecting the number of categories or actual text strings of the category labels. For the *CL* from our example above, the simplest way to achieve this would be to fill them with the numbers 1-9. For now, the actual contents of the *CL*-subsets only matter inasmuch they are distinct from each other and allow reference later on. The number of subsets depends on the number of distinct categories each individual pattern of the structure is supposed to have. This structure can be adapted for other pattern structures, by adapting the number of subsets accordingly.

After having defined the pattern structure, we now also need a set-theoretical representation of the individual patterns. Since we assumed patterns to be statements arranged in certain categories (sets), we now assign said statements to each of the nine subsets for *CL*. We start out by assuming a set that contains all these statements; we define it as follows:

$$S: \{S_1, S_2, \ldots, S_n\} \quad (9)$$

We obtain a pattern by simply assigning certain elements of *S* to each subset of *CL*. For this purpose, we introduce a function *p* from a subset of the set of statements *S* into *CL* into a partition S^p of the sets of statements *S*:

$$S^p = \{Sx_1, \ldots, Sx_n\} = S \quad (10)$$

$$p_i: CL \rightarrow S^p \quad (11)$$

$$P_i = \sum_{CL_j \subset CL} p_i(CL_j) \quad (12)$$

What happens here is that each subset of *CL* (CL_1 to CL_9) gets filled with actual content by having a certain number of statements (i.e., part of the partition of *S*) assigned to it. Partitioning *S* ensures that none of the categories remain empty, i.e., that the pattern is complete. Note, that we use the term 'statement' in the loosest sense, so it can refer not only to full sentences, but also to single words or sequences of words, which are not full sentences, or even images. Therefore, partitioning *S* into clusters of sensible information (= subsets) is a necessary step that should be reflected in the formal analysis. The relation p_i determines the assignment of one subset of S^p to each subset of *CL* (CL_1 to CL_9 in our example). The actual pattern is the sum of all values of p_i returned for all values $CL_i \subset CL$, viz. all categories of the pattern structure (CL_1 to CL_9 in our example). The results is a set of *i* number ordered pairs, which, in our example, might look like the following:

$$P_1 = \{<CL_1, Sx_1>, \ldots, <CL_9, Sx_9>\} \quad (13)$$

Note that the P_1 above is only one possible example out of many. The actual values assigned by p_i are left undefined in this framework, since this depends on the context and proposed content of the individual patterns. The framework

should be flexible and widely applicable, so the actual pattern generation must be left to the domain experts. We can use the pattern relation to generate as many relations p_1 to p_n as we need, and thus to generate n number of *CL*-patterns P_n.

We can apply this analysis to any other similarly structured pattern language *PL* and its subsets PL_i, thus granting us a basic structure of patterns as collections of sets of statements. Of course, simply arranging patterns into sets and subsets does not in itself guarantee that any of these patterns are actually useful or reasonable. But that is also not the purpose of the framework at this stage. What this analysis can tell us is (a) the pattern language or structure *PL* (*CL* in our example) the patterns are generated in, (b) how many statement categories (viz., subsets of *PL*) a successful pattern generated in that language must contain, and (c) which statements can be found in which category, i.e., the patterns themselves (P_i). As we can see, this elementary analysis has already yielded a quite flexible starting framework, via which we can express a pattern structure, partition the information to be transformed into a pattern, and relate that information to the pattern structure. Most importantly, building the framework did not require us to reference the actual contents of *S* or its subsets, meaning that it is – at least so far – applicable independently of its contents or the context it will be used in.

VI. DESCRIPTORS

Even at this elementary level, we can do more than merely put statements into a certain structure and add them to a collection of similar statements. We can also lay the groundwork for the relations between individual patterns of a certain collection. A pattern collection is more than an unstructured assortment of statements and needs some kind of inner structure. At this point, there are three important aspects that an individual *CL*-pattern does not tell the reader at this stage, but which we can capture even at this basic formal level. These are (a) which *other* patterns might be useful or even necessary for a given purpose, (b) exactly at which point during a given task or activity and (c) in which order will they be needed. A design pattern for, e.g., menu depth might sensibly be followed by a pattern for hierarchical structures, and preceded by a pattern for menu types and their suitability. But with only one pattern at hand, one can only guess what else they might need upon being presented with only a single pattern or depend on prior experience. It would be undesirable and arguably defeat the purpose of patterns, if extensive meta-knowledge were necessary to be able to use them successfully. Patterns usually contain references to other patterns as a separate field, though the reliability of this depends on the respective pattern collection. To capture (a), (b), and (c) on a more general level, why we enrich the basic set theoretic framework with specialized descriptor sets, which serve to understand patterns in context with each other. These will allow us to add an additional layer of expressiveness and flexibility to the language.

At its core, a descriptor is nothing more than an ordered set containing several subsets with patterns. By employing ordered sets, we can distinguish its subsets solely by virtue of which position they have within the ordered set. The general idea is to use this property of ordered sets to implicitly add auxiliary information to any given pattern collection, simply by arranging that collection in a certain order. That way, the general structure of a descriptor set needs to be defined only once and one can add additional information to a pattern collection by arranging them in a certain order according to the descriptor. Let us illustrate the basic idea via an example descriptor set. Remember that angle brackets ('<' and '>') denote an ordered set, as opposed to regular sets, which are denoted by curly brackets ('{' and '}').

$$D^E = <\{P_1, P_2, P_3, P_4\}, \{P_5, P_6, P_7\}> \quad (14)$$

Assume that there is a pattern collection that consists of seven patterns. Of these, four are regular patterns and three of them antipatterns. We can now define a descriptor as an ordered pair consisting of exactly two subsets. The first of these subsets contains only patterns, the second contains only antipatterns. By applying this knowledge to D^E, we learn that patterns P_1 to P_4 are regular patterns, and that P_5 to P_6 are antipatterns. We have thus provided an easy way to categorize patterns as regular patterns and antipatterns, that can be applied independent of context, and which is as simple as arranging the patterns in a certain order. Furthermore, we have added information to the pattern collection without having to edit the patterns themselves. (Please note that whether a pattern is a regular one or an antipattern is usually considered essential information and already part of the pattern itself. This is merely an illustrative example that employs two obviously distinctive pattern attributes). In the framework, we will use this structure to make a more general distinction of mandatory vs. optional patterns (see Section VI.B).

Structuring the pattern collection in this way allows for added efficiency when generating new collections, and also facilitates sharing and consolidation of pattern collections. E.g., if one would program a pattern database in this framework, new pattern collections can be categorized by arranging the pattern labels in a certain order, mandated by the predefined descriptor sequence. Similarly, new patterns can be added and enriched with information by simply assigning them to an appropriate subset of a descriptor. One could even consolidate patterns from different sources and/or authors into one collection and categorize them by simply arranging them in a certain order. There are often many pattern collections dealing with similar topics yet the valuable knowledge in these patterns is often difficult to consolidate, simply because pattern approaches vary so much.

Therefore, we view the added advantages gained by adding descriptors as an important quality of the framework and a necessary step towards a pattern framework that facilitates exchange both within and between disciplines. In the following subsections, we will build a standard descriptor set structure in a step-by-step manner. To begin, we postulate a descriptor as an ordered set, which consists of a number of

subsets that contain either individual statements or sets of statements.

A. Targets

One single problem rarely occurs in isolation, but is instead often part of a higher-level problem or occurs while trying to achieve a certain overall goal. These are often nothing more than a single sentence or a few words, but they serve as a good overall indicator about where to find a solution to a particular problem one may have. E.g., a programming pattern might be part of the larger problem of trying to avoid pointer errors in C++. Another example would be Tidwell [12], who structured their design patterns as part of several categories, such as "Organizing the content" or "Showing complex data". One individual pattern can conceivably be part of several such higher-level problems or be used in similar or different context to achieve different goals. This is different from the problem described in a pattern, since a given high-level pattern could very well reference a lower-level problem that addresses a different problem, while both serve the same general purpose. In the following, we will label these high-level problems or overall goals *Targets* (or *T* for short).

Finding, iterating and validating patterns is a lengthy and multi-stage process. Whereas finding a new context a pattern can be used in might be as simple as trying to apply it and succeeding. This is were Targets as separate and standardized entities come in very handy. The Target(s) of a pattern should not be part of the pattern itself, since that would entail having to change and subsequently revalidate a pattern each time a new application possibility for it is discovered. Instead, Targets are separate from the actual patterns, which can be assigned or mapped to them. This allows reusing and reapplying patterns (one of their key aspects) in different contexts without having to modify the patterns themselves each time. Whenever a new application for a certain problem strategy is found, a new Target expressing that application area can be created and the pattern (or several) assigned to it. Due to their general nature and labeling function, Targets are the first entities that will be part of our standardized descriptor structure. This is also one of the reasons why we postulated descriptors as containing either statements or sets of statements. Each pattern in this framework is, per definition, a set of statements. But not everything, which might be a sensible addition to the descriptor, is necessarily a pattern (such as Targets). The first set of statements in a standardized descriptor set is thus always an expression of the Target of a pattern collection. At this point, the descriptor structure looks like:

$$D = <T> \quad (15)$$

'T' is a placeholder for a set containing one or several statements, so a descriptor at this stage could read, e.g., $D^E = <\{S_{37}\}>$ or $D^E = <\{S_{28}, S_{29}\}>$. Of course, a Target without any patterns is rather useless, so we need to add these to the descriptor as well.

B. Mandatory and Optional Patterns

The second subset in the descriptor will contain the actual patterns that are supposed to contribute to the Target expressed in the first set. However, there is one important distinction that we can make at this stage, which consists in separating the patterns into mandatory and optional patterns. Patterns can be part of solutions to higher-level problems and are ideally applicable in similar contexts. It is reasonable to assume that a high-level Target might cover a high-level context and thus a range of several low-level contexts. But not all of these low-level contexts might be similar enough to be interchangeable, thus excluding some patterns depending on which subcontext it is applied to. In addition, solving one problem via a pattern, might (and often will) pose a new problem, for which there are several possible solutions (and thus, several possible patterns).

To illustrate the concept of mandatory and optional patterns via a brief example, assume that we design an interface and want to display items and their contents on screen at once. We decide to take a look at Tidwell's pattern [12] collection and find a pattern titled "Two-Panel Selector", which suits our needs. Following the pattern solution, we divide our interface into two parts; one showing the items, and the other their contents. We then find that our item structure is multi-layered and rather complicated, and that two panels are not sufficient to display it in an adequate fashion. Conveniently, we find a pattern section that contains solutions for displaying complex data. Among these, we find a pattern showing the use of cascading lists and another showing the use of tree-tables. Depending on other considerations (e.g., horizontal space, consistency with the rest of the interface), we will then decide for one of the two solutions, but very likely not both. They are two solutions for a similar (and in our case the same) problem and we are free to choose the one we deem more appropriate for our purpose.

If, however, it is – for whatever reason – impossible or undesirable to separate the patterns into mandatory and optional ones, the set of optional patterns can also simply be left empty. Since the empty set is an abstract entity, these standardized descriptors will always remain compatible with each other. Even when one ore several of their subsets are empty, the *number* of these subsets never changes. Thus, the standard sequence and meaning of each subset is preserved, regardless of whether any of its preceding subsets is empty or not. Again, we further illustrate this via example descriptors:

$$D^E_1 = <\{S_{37}\}, \{P_1, P_2\}, \{P_3\}> \quad (16)$$
$$D^E_2 = <\{S_{38}\}, \{P_1, P_2, P_3\}, \{\}> \quad (17)$$

Both of these example descriptors are of the same structure. They differ in that they have two different targets, and that P_3 is an optional pattern in D^E_1, and a mandatory pattern in D^E_2. They both have the same number of subsets, so if we were to add another set to the descriptor set, we could so without worrying about potentially empty set, since the sequence is preserved.

C. References to other pattern collections or other sources

Patterns are not the only things that can be considered optional when tackling a problem. Other sources and references are often needed as well. Patterns usually contain references to sources they draw from, aside from references to related patterns from the same pattern collection. But there is additional benefit when these references are added at the descriptor level. That benefit is flexibility. The descriptors are not part of the patterns themselves and can be generated at any time, once a pattern collection is available. Thus, any information that can be added by modifying the descriptor does not necessitate modifying the patterns contained in the descriptor. Thus, a seemingly outdated pattern collection can be updated *ex post*, by generating descriptors containing references to material that was not available at the time the patterns were originally generated.

The same can be done with patterns from other pattern collections that handle similar issues. Patterns from other pattern collections can be added to the set of optional patterns. Since the descriptor allows inclusion of any set of statements, this could, in semantical terms, be the full pattern or merely a link to its website or bibliographical reference. Even if the *CL*-structure of both pattern collections were the same, the pattern relations p_i would be different. This means that the "foreign" patterns would not be part of the set of patterns P_i and, thus, are merely sets of statements and easily distinguishable from one's own patterns. The descriptor structure allows consolidation of knowledge from different sources and goes beyond the possibilities gained by referencing only within the patterns themselves.

We show how such references might work via another brief example: Assume that we intend to design a car interface, for which we have our own car interface design pattern collection. While designing, we notice that the interface structure has become very deep and rather difficult to navigate. We now intend to solve this problem by either reducing the menu depth or presenting the information in a more effective way, but cannot find an appropriate pattern in our collection. However, we find such solutions in other, more generic interface design pattern collections. One of these turns out to be particularly to our liking and can be easily applied without any modifications. Both pattern collections were printed and published several years ago, so a revision would not be a trivial task and require substantial effort.

Then, we decide to collect our car design patterns in a database and arrange them via descriptors. To keep the required effort at a minimum, we add the patterns simply as uniquely identifying labels for the original patterns. The resulting database allows us to search for design problems via their Target. We create on descriptor, which has "Designing car interfaces with high menu depth" as its Target and reference the foreign pattern we found in this descriptor. Thus, anybody who faces the same problem and uses the database will know that there is a different pattern collection that provides a solution to a certain subproblem. Such information is normally either included when a pattern is generated or not at all. With the descriptor structure, it is as simple as new descriptor. We can include newly created patterns by inputting them directly or referencing them, as well as draw from knowledge from related fields by referencing other patterns in this way.

By adding the two additional subsets for mandatory and optional patterns and references to the initial descriptor structure, the updated descriptor structure is:

$$D = <T, M, O> \tag{18}$$

D. Pattern Sequence

Finally, problems do not always occur at random, but can appear in a certain sequence. Thus, a solution to one problem might be followed by another solution, dictated by the underlying sequence of problem occurrence. This might be as simple by one problem followed by the other, or it could also reflect a hierarchical structure of high to low-level problem solutions, where depending on how a higher-level problem is solved, different lower-level problems occur. We can find such sequences, e.g., in Breitenbücher et al. [25], who propose a method to organize low-level solutions (so-called *idioms*) in sequence, while taking into consideration the preceding idioms.

To handle sequences at the descriptor-level, we add another subset to the descriptor. The purpose of this set is to put our patterns and other information into a sequence; therefore, we label this additional set *S*, which overlaps with the previously introduced sets *M* and *O*. Since S is supposed to handle only sequences, it would make little sense for something to be part of *S* but not of *M* or *O*. Therefore, we postulate that every element of *S* must also be element of either *M* or *O*. Unlike *T*, *M*, and *O*, *S* is not a regular set but an ordered one. Since the order of their contents does not matter for regular sets, but does matter for ordered sets, arranging the contents as an ordered set is a simple and efficient way to express a sequence.

If we wanted to express that, e.g., P_1 from D^E_1 above would be needed after the solution described in P_2, we could add the sequence set $<P_2, P_1>$ to it and arrive at the following descriptor D^E_3:

$$D^E_3 = <\{S_{37}\}, \{P_1, P_2\}, \{P_3\}, <P_2, P_1>> \tag{19}$$

As we can see, the sequence set contains nothing that is not also in *M* or *O*, and merely puts some (but not all) of the parts of the pattern collection into sequence. We can use the sequence set to not only express linear sequences, but also hierarchical structures. Assume, that we have one high-level pattern P_1 and a number of low-level patterns P_2 to P_7. P_1 proposes three possible solutions to the high-level problems. Depending on which solution is chosen, new low-level problems occur that are described in P_2 to P_4. Each of these solutions is then followed by another set of possible problems, described in P_5 to P_7. We can express this hierarchical structure via the following sequence set S^E_1:

$$S^E_1 = <\{P_1\}, \{P_2, P_3, P_4\}, \{P_5, P_6, P_7\}> \tag{20}$$

The contents of S^E_1 represent a hierarchical structure from high to low. Since the contents of S^E_1 are regular sets, the order within these sets does not matter and they can be considered as being of the same level. We can use the sequence set to specify such a hierarchy even further. If we know which low-level problem leads into which, we could also formulate the alternative descriptor S^E_2:

$$S^E_2 = <\{P_1\}, \{<P_2, P_5>, <P_3, P_7>, <P_4, P_6>\}> \quad (21)$$

In S^E_2, we find P_1 is still the highest-level pattern and can see additionally the sequences between the individual patterns from the lower-level subsets. While it might not be immediately obvious, the three-level structure from S^E_1 is also preserved in S^E_2, since the first element of each ordered pair in S^E_2 is also an element of the middle set in S^E_1.

E. Adding it up

By combining all of the sets T to M, we arrive at the following, final descriptor structure:

$$D = <T, M, O, S> \quad (22)$$

Target, mandatory patterns, optional patterns and references, as well as the sequence set allow for a good amount of expressive possibilities. We will once again illustrate the potential use of the final descriptor structure via a simple example. Assume that we have four patterns, which would help us in conducting a user study in the car. P_1, P_2, and P_3 are patterns to reduce user distraction and part of our own pattern collection. We have also access to another pattern about processing the data gained from the study. This pattern, we label it P^F_1, was generated in a different pattern structure and is, therefore, not part of our patterns P_i. We can now specify which of these patterns we want or need and in which order by introducing a descriptor. To do that, we need to specify the contents of each of its subsets. We further want to express that we definitely need P_1 and P_2, as well as P^F_1 and that the *DL*-pattern will be needed after the *CL*-pattern. We thus arrive at the following example mandatory pattern set M^E:

$$M^E: \{P_1, P_2\} \quad (23)$$

We also know that P_3 has proven useful in several similar cases in the past, but not in all of them, so we consider it as an optional pattern. Having one optional pattern (P_3) and one foreign pattern (P^F_1), gives us the following example set O^E:

$$O^E: \{P_3, P^F_1\} \quad (24)$$

We further know that P_3, should it be needed, is always needed after P_1. P_2, on the other hand, has no fixed position in the sequence, but occurs after P_3 in a few specific cases. P^F_1 is always needed last. This results in the following two sequence sets:

$$S^E_3: <P_1, P_3, P^F_1> \quad (25)$$
$$S^E_4: <P_1, P_3, P_2, P^F_1> \quad (26)$$

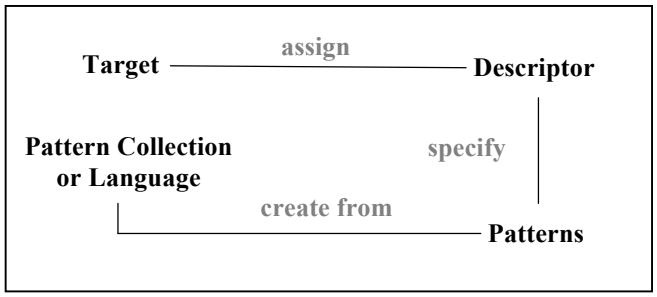


Figure 1. The Pattern Framework – a high-level overview

But how do we now specify which of these sequences is the appropriate one for a given scenario? Since patterns are created for a certain purpose, we need to map each sequence to its most appropriate purpose We can specify this via the Target, which contains the general purpose or overall problem of a collection of patterns. We can now introduce two targets, T^E_1 and T^E_2, with T^E_1 outlining the general high-level problem and T^E_2 specifying the contexts in which P_3 is followed by P_2. These can be any statements; in our example we specify them as the sets $\{S_1\}$ (for T^E_1) and $\{S_1, S_2\}$ (for T^E_2). As a result, we get the following two example descriptors D^E_3 and D^E_4:

$$D^E_3 = <\{S_1\}, \{P_1, P_2\}, \{P_3, P^F_1\}, <P_1, P_3, P^F_1>> \quad (27)$$
$$D^E_4 = <\{S_1, S_2\}, \{P_1, P_2\}, \{P_3, P^F_1\}, <P_1, P_3, P_2, P^F_1>> \quad (28)$$

In addition to being able to specify the relations between patterns from a single pattern language, we are not confined to that single pattern language. Furthermore, we can describe hierarchical and sequential pattern structures from different domains and pattern languages in the same framework.

***D*: <*T*, *M*, *O*, *S*>**

D: Descriptor – contains 4 subsets
(contains 4 subsets, ordered set)

T: Target
(contains a finite set of statements, regular subset of D, may not overlap with M or O)

M: Set of mandatory patterns
(contains a finite set of CL-patterns, regular subset of D, may not overlap with T or O)

O: Set of optional patterns and references
(contains a finite number of sets of statements, regular subset of D, may not overlap with T or M)

S: Sequence set
(contains a finite number of sets of statements, ordered subset of D, every element of S must also be element of M or O)

Figure 2. The Descriptor structure

Furthermore, patterns can be clustered and applied for different purposes in an efficient manner by simply altering the structure of these pattern clusters, and without having to change the patterns directly. By adding one additional layer (targets and descriptors) to what was already available before, we have arrived at a highly modular and flexible pattern framework. Figure 1 provides an overview of the interrelation of pattern languages, patterns, descriptors, and targets. Figure 2 contains a summary of the standard Descriptor structure.

Descriptors can be generated on an as-needed basis, which means that they can be used to categorize the initial pattern collection, as well as update it with additional information, to keep a pattern from becoming outdated quickly without being cumbersome to maintain. Most of the information is contained in the patterns themselves, but we tried to elevate information that might change or be in need of frequent updates to the descriptor-level. This makes expanding and updating pattern collections easier, since most of these changes will require either changing or generating descriptors, which are nothing more than strings in a predefined sequence. This approach has the additional advantage that descriptors can be used to consolidate knowledge from different sources, since the descriptor structure is based solely on the set theoretic framework and, therefore, not bound to any particular field or context. We have thus provided a framework that is formally sound, based on only elementary mathematical principles, flexible regarding its content, and with additional means of referencing and consolidation via the descriptor structure. In order to actually use the framework for structuring patterns, however, one still requires a means to collect and structure data regarding working solutions. We provide some general recommendations on how to do that in the following section.

VII. ADDING THE CONTENT

We mentioned in Section V that the pattern relation p_i is left undefined in the framework and is only specified insofar as it is a relation from CL into S^p. Similarly, the specifics of the partitions in S^p are undefined since these vary depending on pattern structure and content. This ensures flexibility of the framework, but it also means that actually generating patterns cannot be done with only the framework itself. The framework ensures consistency and easier means of referencing and consolidation across domains, disciplines, and pattern languages. The pattern framework is intended for a wide audience, which includes those who are not yet familiar with pattern approaches but would want to apply them in their field. Therefore, we use this section to provide a consolidated overview of some of the literature regarding pattern generation, to supplement the formal framework. This is not meant to be a comprehensive summary, but a general aid to generate meaningful patterns in this particular framework. In the following sections, we mainly draw from Meszaros and Doble's [9] pattern language for pattern writing, Winn and Calder's [10] pattern language for pattern language structure, and some of Borchers' [7] considerations regarding pattern generation.

A. Defining the structure

One of the primary steps when beginning to build a pattern language – and elementary to partitioning the statements that will later constitute the individual patterns – is to define the pattern structure. By that, we do not mean the relations or hierarchy between patterns, but rather the number of sub-categories or fields of each pattern. Pattern structures exist in a wide variety of granularity. Tidwell's [12] pattern structure is minimalistic but effective, with only six subcategories (what, use when, why, how, examples, in other libraries), whereas the structure introduced by Gamma et al. [4] propose 13 subcategories for each of their patterns. The exact number of subcategories should be decided on individual needs, preferences, and also available resources (more subcategories = more complicated and longer pattern mining process). However, there are a few basic subcategories, which each pattern structure should contain. We present these, together with the reasons why we consider them to be essential, in the following.

Name: Patterns should be uniquely identifiable, so that they can be referred to and structured with regard to other patterns. Therefore, each pattern should have a unique name that clearly distinguishes it from other patterns. It is furthermore helpful if that name is not obtuse or even presents an image of the suggested solution in the reader's mind (Meszaros and Doble refer to this as an "evocative pattern name" [9]).

Problem: One of the major distinguishing features of patterns is their problem-centric nature. If the pattern does not present a solution to a (reoccurring) problem, then it only provides general guidance and serves the same purpose as a guideline, but without the comprehensive character a guideline usually provides. Therefore, a separate description of the problem is considered essential for a successful pattern.

Context and/or Forces: Patterns contain proven, working solutions, which means that these solutions solved the problems in particular cases. Therefore, understanding and documenting this context is elementary for being able to decide whether a particular solution is suitable for a different (even when similar) context. Forces are the aspects of the context that the solution is supposed to optimize. They are important, but not always considered as separate pattern subcategories (e.g., [1], [12]). Therefore, we only consider one of them as essential – unlike Borchers [7] and van Velie [5]. The bottom line is that each pattern should, at the very minimum, contain some kind of description of its context as a separate entry – whether it be *context* or *forces* (or both).

Solution: A seemingly obvious point that is never the less worth pointing out. Each pattern should contain a description of the actual solution as a separate entry. This is not the same as a simple screenshot of a working example, but rather a detailed textual description of the steps taken to solve the problem in its particular context.

Examples: Since the solution described by the pattern is supposed to be a proven one, concrete examples (preferably more than one) should be provided to show the end result of the implemented solution. These examples are closely

related to, but not the same as, the solution. They help to put the general solution into more practical terms and link the solution to its context. In the case of several implementation examples being available, they can also aid the designer in identifying essential commonalities between application contexts. This can be an additional aid, when a designer is not sure whether a pattern would be suitable for their particular context.

A successful pattern structure can have as many pattern subcategories as needed, though the ones listed above should be considered a reasonable minimum for any pattern structure. The minimum requirements we presented here are very similar to those given by Gamma et al. [4] (pattern name, problem, solution, consequences), but with slight extensions and modifications for wider applicability. We also decided to not include an implementation's consequences as a necessary component, since that might be a confusing concept for patterns outside of areas in which consequences can be traced more easily (such as in software code, where changes and the parts they affect can be more or less fully described).

B. Mining and Iteration

In order to generate meaningful patterns, the solutions contained therein need to be discovered first. Pattern generation is a difficult and lengthy process, which usually occurs in several phases. Köhne [25] describes the pattern generation process as consisting of the following 8 stages: pattern mining – pattern writing – shepherding – writer's workshop – review by pattern author – collection of pattern in repository – peer review – pattern book publication. While this is a good overall summary of how pattern generation or finding occurs, Pattern creation does not always follow these exact steps in reality. There is no single accepted method or process for pattern generation, but there are several useful recommendations for generating successful patterns by Borchers [7], Martin [15], Vlissides [17], and others. In the following, we present what we consider the bare minimum of what a pattern generation process should entail.

The first step in generating a pattern is recognizing the problem and its reoccurring nature. There is no standard procedure for this and Appleton [18] even notes that the best way to learn how to recognize patterns is to learn from others who were able to do so successfully. This is why pattern generation should happen in several stages. Anyone, who has worked in a certain field for some time, should be able to eventually spot problems that have manifested themselves over and over in the past. They might also be able to recognize certain regularities in the solutions that were employed to solve the problem in all its past occurrences. To go from this initial pattern assessment to a complete pattern, examination and iteration should happen in several steps and by several people, so that the essence of the solution can be extracted and adequately described. Furthermore, reexamination and iteration should be done by several individuals. These pattern iterators will then rework the patterns to suit their readability requirements, i.e., the resulting pattern will automatically be written and formatted for easier readability for a wider audience. Even if the pattern started out as a simple assumption about a potential solution, at the end, the pattern contains the know-how of all its iterators and a quantitative component that complements the pattern content. After all, if multiple experts came to similar conclusions about a problem and its solution, then this lends support to the assumption that the solution is indeed a working one and the problem a reoccurring one. Thus, it can be possible even for people who are inexperienced in pattern generation to come up with successful patterns.

Therefore, the most important steps any successful pattern generation process should contain are (a) *problem identification* to define the elementary parts, context, and eventually the solution; (b) *structuration* to guarantee a uniform format, good readability, and completeness of patterns with the same structure; and (c) *reflection and feedback* to examine whether the solution is a working one and ensure sufficient detail of its explanation to allow easy application.

C. Piecemeal Growth

This point is based on Winn and Calder's [10] suggestion by the same name. They suggest, "if new structure needs to be added to the system, then add it gradually, piece by piece, evaluating the effect of the change on the whole." In their work, Winn and Calder have applied this to systems (software, architectural, biological), as well as pattern languages. In this paper, we adapt their ideas only for the generation of pattern languages.

Building a full pattern language is a lengthy process, which begins with a few patterns. As more solutions are discovered, more patterns can be created, which culminates in a full pattern language, once a certain number and level of comprehensiveness of patterns is reached. This means that new solutions and, therefore, new patterns must be considered in light of already existing solutions. It is possible that a new solution is incompatible with an already established solution, where both problems usually occur together. In such a case, parameters must be provided that allow deciding when one or the other solution should be applied. Similarly, a newly introduced solution might be superior to a previous solution, rendering its respective pattern obsolete. This must be reflected in the pattern language, as they would otherwise seem like equally effective solutions to the same problem. Therefore, changes and additions to any existing patterns should occur in small steps, while re-evaluating the existing patterns in light of these new additions.

In terms of the pattern framework, this means that newly generated patterns should ideally entail review and potential modification of descriptors. Since descriptors allow mapping patterns to overall goals, modifications to the existing patterns themselves should seldom be necessary. An initial pattern collection might only have a single descriptor, since the patterns are likely to be generated with one overall goal or problem in mind. However, it is very possible that a new pattern presents a solution that often, but not always, occurs with other problems for which patterns are available. In such cases it is recommended to create to separate descriptors that

cover both cases – those, in which both problems occur and those, in which they do not. The same is true for conflicting patterns. These can be put into different descriptors, thus making these conflicts visible without a need for modification of any of the patterns themselves. In the case of outdated patterns, these can simply be left as they are, but not made part of any descriptor. Therefore, they are still available for reference purposes, but not part of any recommended set of solutions.

Cases, such as the ones described above, which necessitate a restructuring of both new and existing patterns, can happen at any stage in the pattern language development process. However, the additional pattern does not necessarily entail a new descriptor. It could simply be added to an existing one or prompt the creation of several new ones, all depending on the individual case. Therefore, the growth of a pattern language's complexity cannot be considered linear in regard to the number of patterns it contains.

The development of a pattern language can be seen as an organic process, where changes and additions can have wide-ranging consequences. Therefore, such changes and additions should happen in small steps, followed by a reexamination of the pattern collection. In the framework, this reexamination should almost always happen at the descriptor-level.

D. Cross Linkage

This point ties in with the previous one and is, once again, strongly grounded in Winn and Calder's [10] suggestion by the same name. They state, "if the system structure is complex, then overlap and use cross linkages to capture complexity." The general idea is that linear or linear-hierarchical structures cannot be a catch-all for complex structures. A pattern structure should allow cross-linking and overlaps between its elements, so that it can support complex structures.

In the previous section, we explained that even a single new pattern could potentially entail fundamental changes to the overall pattern collection. In the framework, this can manifest itself as the creation several new descriptors or the vanishing of older, outdated descriptors. In order for this flexibility to be possible, overlaps and links between the descriptors must be possible, which is the case for the framework due to its basis in set theory.

Different descriptors can largely have the same contents, with only minor differences, to satisfy different Targets. For example, two descriptors might differ in one only containing one more pattern than the other, thus dealing with a special case of the other's, more general, Target. They might even be identical regarding their elements, but with different sequence sets. One of these descriptors could then serve as a solution to a hierarchical occurrence of the problem, the other to a differently structured overall problem. Patterns are supposed to be reused in similar contexts; the descriptors, therefore, support that reuse and allow multiple occurrences of the same pattern and overlaps between descriptor contents. To adequately support the nonlinear growth in pattern language complexity when new patterns are added, it is important to generate as many new descriptors as necessary once new links between patterns or Target hierarchies are discovered.

It is not uncommon that a pattern language would start out as a neatly organized string of patterns that all serve one universal goal. As the language's complexity grows, so should its level of detail. A neatly organized descriptor variety helps structuring and reapplication of patterns for different contexts. It is also an invaluable aid for efficient and quick searching and finding of solutions to particular problems. A designer or practitioner will likely not need the whole pattern language for any given task, but also not know which individual patterns they do need, unless they read through the whole pattern catalogue. By employing the proposed method, only the individual descriptors need to be read to identify, whether a patter cluster that provides a solution to a certain goal or not. Once one is found, the reader is led through all relevant patterns, their links, and in the proper sequence via the descriptor's structure.

VIII. The Framework applied – Car User Experience Patterns

An actual pattern collection usually takes either the form of a (often online) pattern database or a printed volume. The framework was constructed mainly with databases in mind, since the added flexibility by using descriptors is easier to realize when existing input can be added to (which is difficult to do with published paper collections). In addition, the sets of statements that make up each pattern category are directly translatable into data fields and the descriptors can then point to these data. While the framework loses some of these advantages when applied to a paper-based pattern collection instead of a database, it is still feasible to use it for that purpose. In this section, we present an example of a paper-based User Experience design pattern collection, which was structured using the universal pattern framework.

The pattern collection consisted of 16 individual patterns. All of them were about design problems in the car with the aim to reduce mental workload while interacting with the interface. The actual pattern finding process is described in detail in [26]. The resulting patterns all followed the same structure, which consisted of nine categories of statements (Name, Intent, Topics, Problem, Scenario, Solution, Examples, Keywords, Sources). Since the descriptor still enables structuring towards overall goal and regarding pattern sequence and status (mandatory vs. optional), we created one descriptor to serve as an index for the whole pattern collection.

The overall goal of every pattern was to provide design solutions that reduce mental workload, so the appropriate Target became *UX Factor: Reduction of mental workload caused by distraction in the car*. 'UX Factor' was added since this is one of several factors that are postulated to influence UX and to distinguish these from later patterns that address different influences on UX. The patterns were findings from scientific works, supplemented with implementation examples, and iterated in collaboration with industry stakeholders. Due to this somewhat nonstandard

approach, there were some patterns that had more implementation examples and more straightforward instructions on how to put their respective solutions into practice. However, others provided less such examples and were perceived to be more suited for more experienced designers. Therefore, this second type of patterns was considered optional and only for those who have the necessary skills to put the proposed solutions into practice.

Thus, we described these two sets of patterns via the descriptor's sets for mandatory and optional patterns and references. Finally, there was one pattern that relied on another pattern from the same collection. Using the solution in the first pattern could sometimes create an additional problem, which the second pattern would help to solve. But it would have been misleading to imply a necessary connection and write one single pattern for both problems, since they occur together only sometimes, but not always. In order to adequately represent this relation, the two patterns were put into the sequence set to indicate that reading the first should always entail reading the second one afterwards. In the text, we indicated this with one sentence between the patterns explaining the possible link. With all this taken into account, the resulting descriptor looked as follows:

$$D_1: <T_1, \{P_1, \ldots, P_{11}\}, \{P_{12}, \ldots, P_{16}\}, <P_5, P_6>> \quad (29)$$

This was then transformed into an index. The Target served as the overall headline, patterns 1 to 11 and 12 to 16 were put into separate subsections, and patterns 5 and 6 were put into sequence and linked explicitly with additional text between the two patterns. By using the framework approach, we were able to easily structure the pattern collection in a meaningful way, even though the framework contains no information that would be specific to the car or UX-domains. Moreover, the thusly-structured pattern collection can still be put into a database, without a need for

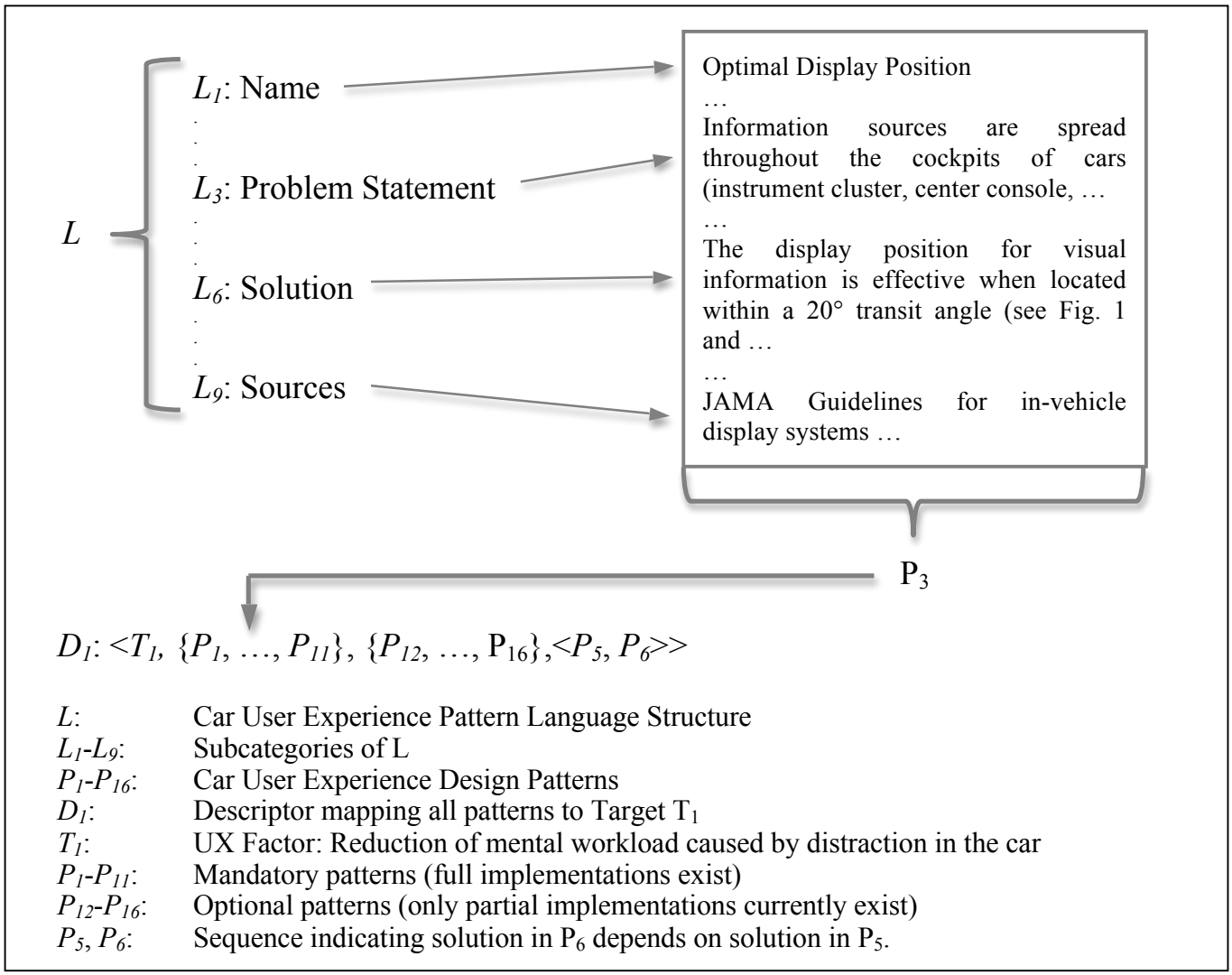


Figure 3. Car User Experience Pattern Descriptor and Pattern Example

any substantial restructuring work, since the set- and descriptor structures are consistent among all pattern collections that are based on the framework.

This collection of 16 patterns is part of a prospected larger collection spanning two more UX factors in addition to the first one. At the writing of this paper, the pattern finding process for these additional two UX factors had not yet been completed. Nevertheless, the end result will be a collection of several patterns, which are mapped to the three different UX factors and structured internally via the framework's proposed descriptor structure. Figure 3 provides an additional overview of the descriptor structure, along with an example for how an individual pattern relates to the pattern structure in the framework. To reiterate, an individual pattern is a set, which contains several subsets of statements. A pattern collection or language consists of several such sets. Descriptors are separate sets, which are used to map individual patterns to an overall goal, and can thus be used to structure the pattern set as well as map lower- to higher-level patterns. There is no a priori limit to the amount of descriptors that can be created for any given pattern collection. The number of descriptors depends on the amount of goals, which are identified and/or deemed necessary for any given purpose.

IX. DISCUSSION

Next, we discuss the proposed framework, address benefits, possible shortcomings, and future work potentials.

A. Benefits of using the framework

The framework provides a flexible basis purely by virtue of its formal features. The basic and uniform structure enables any adequately structured set of statements to be considered a potential pattern, so as long as an area or discipline can satisfy this minimal requirement, it can use the framework to structure its patterns. This general applicability also means that the framework cannot serve as a suitable means to verify a pattern's (or structure's) validity or soundness on its own. What the framework offers is a consistent basis, which the individual disciplines can build upon. Pattern languages can be described as partitioned sets of statements within the framework. As long as the structure of a certain pattern collection is known, its individual patterns and sub-parts of patterns can be referenced within the framework by referencing the appropriate set or sub-set.

Thanks to the descriptor structure, linkage between patterns and pattern languages within the framework is possible, even between patterns from different disciplines. In such a case, the differences in pattern structure must be known and appropriately modeled in the framework, since it is very likely that their structures do not consist of the same sub-categories. Mapping patterns to overall Targets can reduce redundancy and allows mapping of lower-level patterns to higher-level goals. The standard structure of Descriptors allows structuring patterns with regard to priority (mandatory vs. optional) as well as sequences of problems or their solutions. Finally, all these features are available on the very basic framework-level, and are thus not dependent on any particularities of the actual pattern content or the discipline they belong to. Thus, the initial goal of the framework not being bound to any individual discipline or domain is achieved.

B. The set-theoretic basis and its multi-domain suitability

As initially stated, the framework is intended as a basis for patterns as a general knowledge transfer tool, suitable for a multitude of disciplines and domains. However, employing mathematical methods might seem to limit the framework to only those disciplines already familiar with such methods, which is why we briefly discuss the need for this mathematical basis and its consequences for applications of the framework. The framework was developed with databases, as well as paper-based pattern collections in mind. Therefore, a suitable framework should fulfill the minimum requirements of consistency and division of information into separate categories or data fields. This ensures that any pattern from such a framework can be used as input for a database, by treating the pattern subcategories as datafields in the database. By keeping that structure the same for both database and paper-based pattern collections, compatibility and consistency between the two is ensured. This also permits any paper-based collection built in this framework to be incorporated into a database of the same format.

The formulae in Sections V and VI are accompanied by explanations, so that the purpose of the theoretical basis can be understood without necessarily having to understand the methods themselves. Thus, the framework does not require knowledge of mathematics or formal methods to be applied, as long as the separation of patterns into statement categories and the meaning of the descriptor contents are understood by the reader. Such an application of the framework would likely result in a well-structured paper pattern collection, like the one shown in our example in Section VIII. However, as example also showed, a paper-based collection loses some of the framework's advantages. This issue is inherent to the medium, as it is generally difficult to update or crosslink published volumes (short of releasing updated reprints). We do not think that there is any framework that could solve this fundamental issue, so the minimum requirement of handling databases must be fulfilled by anyone who intends to apply the framework to its full extent.

The set theory employed in this framework is elementary and based on conventional (Boolean) logic. The reason for this is, once again, the desire to keep the framework as easy to understand and handle as possible. But furthermore, we believe that for achieving the goals outlined in Section I, conventional elementary set theory is absolutely sufficient, as we merely arrange statements in sets and a statement is then either present in a given set or it is not. There are no degrees involved here that would warrant employing fuzzy operations or sets. The same goes for other extensions to conventional logic and set theory: unless they are needed, they would only complicate matters without adding any tangible benefit (and since they are often supersets of conventional set theory, the framework could still be extended on an as-needed-basis in special cases). A more complicated underbelly would probably not matter for the

average reader with an IT-background and who is already familiar with logic to some degree. But for those with different backgrounds, it might create an additional hurdle that we would rather avoid. The framework is rather simple on a formal level but it achieves what it was meant to do just as well. In this regard, we see the framework to strike the best possible balance between necessary skill level of the user and application possibilities.

C. Finding patterns and descriptors – it's not that easy

Putting patterns into a meaningful structure is only one step in any pattern finding process, although a rather important one. The purpose of this paper was to provide a *basis* for patterns as a universal tool, and not a complete guide for discipline-independent pattern finding. Nevertheless, if we want the framework to be successful, then it should ultimately be applied in areas, in which there have been no (or few) pattern approaches in the past. In such areas, simply providing a framework without any guidance on how to actually *find* patterns would arguably be of little use. Therefore, we included a number of recommendations based on existing pattern approaches in Section VI. We consider these recommendations elementary enough to be sensible for any pattern collection and, therefore, a suitable supplement to the framework. On the other hand, the elementary and general nature of these recommendations also means that they are, at best, necessary (but not sufficient) conditions for successful pattern finding. We acknowledge that the recommendations given in Section VI constitute a sensible starting point but not a complete pattern finding guideline, and that more work on pattern finding (both within and across disciplines) is needed.

D. Tool support

The framework is, in its current state, not supported by a tool or any other automated means that could aid the user in finding patterns or creating a pattern language. The framework provides a basis that is consistent among disciplines but most of the necessary legwork still has to be done by the individuals themselves. This is not something that cannot be fully eliminated, but a completely unassisted framework is a lot less accessible than it could be, esspecially considering our aim of domain-independent applicability.

There are specialized tools that can aid solution finding in certain contexts; the EXPLAINER tool by Redmiles et al. [27][28] is one such tool. Tools like this one might be reusable in other disciplines as well, but it can be expected that full tool support from pattern finding to arrangement in a language, can probably not be handled on a universal level by one single tool. However, since the basic framework is essentially a means to structure statements and set them in relation to each other, there is no reason why it shouldn't be possible to simply provide a database input mask that assists with the most common operations (defining number of category-subsets, labels, adding the statements, defining descriptors with predefined subsets, etc.). This is something that would greatly aid users in applying the framework and we hope to be able to provide such an aid further down the line.

E. Wider application

In Section VIII, we provided an example of an actual application of the framework in practice. The example was for a paper-based pattern collection, which illustrated how the descriptor structure can be used for meaningful categorization within a pattern collection with relatively little effort. Overall, the example might seem rather unspectacular, especially since it only resulted in the creation of one single descriptor. What we did not show was an actual pattern database that makes full use of all of the framework's advantages (most importantly, multiple descriptors for overlapping pattern sets and reference to sources or patterns from outside). We intend to use the framework for many more future pattern collections (including databases), so that more application examples will eventually become available. At this point, the framework is still very new and we do not have a complete database that would be suitable for demonstration purposes. However, we do think that the framework is outlined in sufficient detail in this paper to allow successful application at this stage and we encourage the community to use (and criticize) the framework, as only actual use can really show it suitability (or lack thereof).

X. CONCLUSION

In this paper, we have provided a formal framework that supports finding and structuring patterns independent of their domain, field or discipline, supplemented with information on how to generate actual content (i.e., finding patterns) and gave an example of an application of the framework in practice.

In our framework, patterns are separate from descriptors, which are themselves separate from their targets. This means, that patterns can be generated as usual and assigned on an as-needed basis. For the pattern user, this means that they do not have to scour vast databases of patterns for those they might need. All they need is to have a look at the descriptor(s) that is/are assigned to the target they have in mind. For the pattern provider, there is also the added advantage existing pattern databases can be expanded with descriptors, which help make them more usable and reduce the amount of domain experience and previous knowledge required in order to employ patterns successfully. The example we provided in Section VIII is one such case. The paper version can be made into a database using the same structure and format. Additional descriptors and/or patterns can then be added and the collection expanded as needed.

Descriptors can function similarly to references contained in the patterns themselves (as suggested by Borchers [7]), but enable additional or alternative references to other patterns at any time, since they are not actual parts of a pattern. This means that descriptors can be used to

describe virtually any pattern set, regardless of which domain(s) its patterns came from or when the pattern was created. Not only it is possible to capture the hierarchical order of existing pattern languages via descriptors, but also reference patterns from other languages that might fit a certain purpose. This means that the framework is not tied to a single pattern language or even a single domain and permits references to patterns from multiple pattern languages. The framework still needs to be adopted and used on a wider scale, in order to prove its suitability in practice. Nevertheless, due to its general basis and viability for both pattern databases and paper-based pattern collections, we consider it an appropriate basis for patterns as a domain independent knowledge transfer tool. We will use the framework as a basis for our future pattern collections (including a pattern database implementation) and further iterate the framework, as new insights from such use cases are gained.

ACKNOWLEDGMENT

We gratefully acknowledge the financial support by the Austrian Federal Ministry of Economy, Family and Youth and the National Foundation for Research, Technology and Development (Christian Doppler Laboratory for „Contextual Interfaces").

REFERENCES

[1] A. Mirnig and M. Tscheligi, "Building a General Pattern Framework via Set Theory: Towards a Universal Pattern Approach," The Sixth International Conference on Pervasive Patterns and Applications (PATTERNS 2014) IARIA, Venice, Italy, May 2014, pp. 8-11.

[2] C. Alexander, The Timeless Way of Building, New York: Oxford University Press, 1979.

[3] C. Alexander, S. Ishikawa, M. Silverstein, M. Jacobson, I. Fiksdahl-King, and S. Angel, A Pattern Language: Towns, Buildings, Construction, Oxford: University Press, 1979.

[4] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design patterns: elements of reusable object-oriented software, Boston: Addison-Wesley Professional, 1995.

[5] M. van Velie and G. C. van der Veer "Pattern Languages in Interaction Design: Structure and Organisation," Proc. Ninth Int. Conf. on Human-Computer Interaction, IOS Press, IFIP, Zürich, 2003, pp. 527-534.

[6] D. May and P. Taylor, "Knowledge management with patterns," Commun. ACM 46, 7, July 2003, pp. 94-99, DOI=10.1145/792704.792705, retrieved: May 2015.

[7] J. Borchers, A pattern approach to interaction design, New York: John Wiley & Sons, 2001.

[8] A. Krischkowsky, D. Wurhofer, N. Perterer, and M. Tscheligi, "Developing Patterns Step-by-Step: A Pattern Generation Guidance for HCI Researchers," Proc. PATTERNS 2013, The Fifth International Conferences on Pervasive Patterns and Applications, ThinkMind Digital Library, Valencia, Spain, May 2013, pp. 66–72.

[9] G. Meszaros and J. Doble, "A pattern language for pattern writing," Pattern languages of program design 3, Robert C. Martin, Dirk Riehle, and Frank Buschmann (Eds.). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, November 1997, pp. 529-574.

[10] T. Winn and P. Calder, "A pattern language for pattern language structure," Proc. 2002 Conf. on Pattern Languages of Programs - Volume 13 (CRPIT '02), James Noble (Ed.), Volume 13. Australian Computer Society, Inc., Darlinghurst, Australia, June 2003, pp. 45-58.

[11] K. Devlin, The Joy of Sets: fundamentals of contemporary set theory, 2nd ed., Springer, 1993.

[12] J. Tidwell, "Designing Interfaces : Patterns for Effective Interaction Design," O'Reilly Media, Inc., 2005.

[13] A. Dearden and J. Finlay, "Pattern Languages in HCI: A Critical Review," HCI, Volume 21, January 2006, pp. 49-102.

[14] S. Günther and T. Cleenewerck, "Design principles for internal domain-specific languages: a pattern catalog illustrated by Ruby," Proc. 17th Conf. on Pattern Languages of Programs (PLOP '10). ACM, New York, NY, USA, Article 3, pp. 1-35, DOI=10.1145/2493288.2493291, retrieved: April, 2014.

[15] D. Martin, T. Rodden, M. Rouncefield, I.Sommerville, and S. Viller, "Finding Patterns in the Fieldwork," Proc. Seventh European Conf. on Computer-Supported Cooperative Work, Bonn, Germany, September 2001, pp. 39-58.

[16] M. Falkenthal, J. Barzen, U. Breitenbücher, C. Fehling, and F. Leymann, "Efficient Pattern Application: Validating the Concept of Solution Implementations in Different Domains", International Journal on Advances in Software, issn 1942-2628, vol. 7, no. 3 & 4, 2014, pp. 710-726, http://www.iariajournals.org/software/, retrieved: May 2015.

[17] J. Vlissides, Pattern Hatching: Design Patterns Applied (Software Patterns Series), Addisoin-Wesley, 1998.

[18] B. Appleton, Patterns and Software: Essential Concepts and Terminology, February 2000 http://www.bradapp.com/docs/patterns-intro.html, retrieved February 2015.

[19] E. Blomqvist, K. Sandkuhl, "Patterns in Ontology Engineering: Classification of Ontology Patterns," Proc. Seventh Int. Conf. on Enterprise Information Systems (ICEIS 2005), Miami, USA, May 2005, pp. 413-416, retrieved: May 2015.

[20] R. A. Falbo, G. Guizzardi, A. Gangemi, V. Presutti, "Ontology Patterns: Clarifying Concepts and Terminology," Proc. 4th Workshop on Ontology and Semantic Web Patterns (WOP 2013), CEUR-WS, Sydney, Australia, 2013, retrieved: May 2015.

[21] M. Poveda-Villalón, M.C. Suárez-Figueroa, A. Gómez-Pérez, "Reusing Ontology Design Patterns in a Context Ontology Network," Proc. Second Workshop on Ontology Patterns (WOP 2010), CEUR-WS, Shanghai, China, 2010, pp. 35-49, retrieved: May 2015.

[22] O. Noppens and T. Liebig, "Ontology Patterns and Beyond – Towards a Universal Pattern Language, " Proc. Workshop on Ontology Patterns (WOP 2009), CEUR-WS, Washington D.C., USA, 2009, pp. 179-186, retrieved: May 2015.

[23] A. Gangemi and V. Presutti, "Ontology Design Patterns," Handbook on Ontologies, Second ed., S. Staab, R. Studer (Eds.), Springer, 2009, pp. 221-243.

[24] E. Blomqvist, A. Gangemi, V. Presutti, "Experiments on Pattern-based Ontology Design," Proc. 5th Int. Conf. on Knowledge Capture (K-CAP 2009), September 2009, Redondo Beach, California, USA, pp.41-48, retrieved: May 2015.

[25] U. Breitenbücher, T. Binz, O. Kopp, and F. Leymann, "Automating cloud application management using management idioms," Proceedings of the Sixth International Conference on Pervasive Patterns and Applications (PATTERNS), pp. 60–69, May 2014.

[26] A. Mirnig, A. Meschtscherjakov, N. Perterer, A. Krischkowsky, D. Wurhofer, E. Beck, A. Laminger, and M. Tscheligi, "Finding User Experience Patterns Combining Scientific and Industry Knowledge: An Inclusive Pattern Approach," The Seventh International Conference on

Pervasive Patterns and Applications (PATTERNS 2015) IARIA, Nice, France, March 2015.

[27] D. F. Redmiles, "Reducing the Variability of Programmers' Performance Through Explained Examples," Proc. INTERCHI '93 Conf. on Human Factors in Computing Systems, IOS Press Amsterdam,Amsterdam, The Netherlands, 1993, pp. 67-73.

[28] C. Rathke, D. F. Redmiles, "Improving the Explanatory Power of Examples by a Multiple Perspectives Representation," Proc. 1994 East-West Conf. on Computer Technologies in Education (EW-ED '94), Crimea, Ukraine, September 1994, pp. 195-200.

An Easy and Efficient Grammar Authoring Tool for Understanding Spoken Languages

A Novel Approach to Develop a Spoken Language Understanding Grammar for Inflective Languages

Antonio Rosario Intilisano, Salvatore Michele Biondi, Raffaele Di Natale and Vincenzo Catania
Dipartimento di Ingegneria Elettrica Elettronica e Informatica
University of Catania
Catania, Italy
aintilis@dieei.unict.it, salvo.biondi@dieei.unict.it, raffaele.dinatale@dieei.unict.it, vincenzo.catania@dieei.unict.it

Ylenia Cilano
A-Tono Technology s.r.l.
Catania, Italy
ylenia.cilano@a-tono.net

***Abstract*— In a Spoken Dialog System, the Spoken Language Understanding component recognizes words that were previously included in its grammar. The development of a grammar is a time-consuming and error-prone process, especially for the inflectional or Neo-Latin languages. In fact, the developer must include manually all the existing inflected forms of a word. Generally, a regular software developer does not combine linguistic and engineering expertise in spoken language understanding. For this reason, we developed a tool that produces a grammar for different languages, in particular for Romance languages, for which grammar definition is long and hard to manage. This paper describes a solution to facilitate the development of speech-enabled applications and introduces a grammar authoring tool that enables regular software developers with little speech/linguistic background to rapidly create quality semantic grammars for spoken language understanding.**

Keywords- Spoken Language Understanding; Natural Language Understanding; Spoken Dialog System; Grammar Definition.

I. INTRODUCTION

To build a Spoken Language Application in a specific user language, the developer has to design and develop a knowledge base called grammar for Spoken Language Understanding (SLU). The development of a grammar can be greatly accelerated by using a corpus describing the application or a tool that automatically extends grammar coverage [1]. However, the development of such a corpus is a slow and expensive process [2]. In SLU research domain-specific semantic grammars are manually developed for spoken language applications. Semantic grammars are used by robust understanding technologies [3,4] to map input utterances to the corresponding semantic representations. Manual development of a domain-specific grammar is time-consuming, error-prone and requires a significant amount of expertise. It is difficult to write a rule-set that has a good coverage of real data without making it intractable [5]. Writing domain-specific grammars is a major obstacle to a typical application developer. This specialization often does not cover any unspecified data and it often results in ambiguities [6]. These difficulties are further accentuated if a regular software developer does not know the desired user-language that the spoken dialog system (SDS) uses. A further level of abstraction, especially for the Latin languages is necessary.

To facilitate the development of speech-enabled applications, it is necessary to have a grammar authoring editor that enables regular software developers with little speech/linguistic background to rapidly create quality semantic grammars for SLU [7]. More precisely, the purpose of this paper is to ease the development of a CMU Phoenix Grammar [8], a SLU parser of the Olympus Framework. This is accomplished by introducing an intermediate grammar that helps generating a simpler, reusable, and more compact grammar. The development process allows obtaining large amounts of grammar contents starting from a few rows of the new grammar that we are introducing. The grammar has a greater coverage than the standard grammar developed by a regular software developer. In addition, it is possible to write this grammar in the English language and our tool creates the grammar in the SDS user language. Therefore, we are developing a standard grammar that produces a multiple-language support to an application SDS in a simple way. The effort to build the corpus is reduced by the ability of our tools to automatically extend the coverage of the grammar. It currently supports the generation of a grammar for the Italian language, but the method can be applied to all the Romance or Neo-Latin languages.

In order to test the validity of our solution, a specified grammar editor has been developed. It permits the automatic conversion of the new grammar format to the CMU Phoenix grammar [9]. The purpose of this tool is to increase developer productivity; experimental results show that it also improves the coverage of the final Phoenix grammar.

This paper is organized as follows: Section II describes the behavior of SDSs. Section III explains the features of the

Romance languages with particular regard to the Italian. Section IV introduces Olympus, which is a framework for implementing an SDS, and its grammar parser, Phoenix, that use a particular grammar format. The proposed grammar format method is presented in Section V. The two following sections introduce the grammar generator that takes as input the new grammar format; its components are a Morphological Generator for the Italian language (Section VI) and a grammar editor (Section VII). Section VIII shows an example. Finally, in Section IX, we draw conclusions.

II. SPOKEN DIALOG SYSTEMS

A SDS is a computer agent that interacts with people by understanding spoken language. Nowadays, the SDSs market is a big slice of the human-computer interaction field. Many projects, open source and not, have been developed by several universities and companies. Many of these projects have been integrated into commercial technology. The first generation of SDSs was able only to recognize short dialogs or sometimes only single words. Specifically, each single word was bound to a specific functionality and there was no such a thing as a complete dialogue between system and user. The evolution of technologies and software architectures makes it possible to dialogue with the spoken systems and to perform actions that are the results of a dialog composed by different consequent interactions. Now, Spoken dialogue technology allows various interactive applications to be built and used for practical purposes and research focuses on issues that aim to increase the system's communicative competence.

A. How SDS works

The user starts a dialog as a response to the opening of a prompt from the system. The user utterance is automatically transcribed by the Automatic Speech Recognition (ASR) component. The ASR takes a speech signal as an input and produces its transcription in textual format. The SLU module takes the output of the ASR module and generates a meaning representation. Based on the interpretation coming from the SLU module, the Dialog Manager (DM) selects the next dialog turn, this is converted into a natural language sentence by the Natural Language Generation (NLG) module. Finally, the Text-To-Speech (TTS) module synthesizes the generated sentence as a speech signal, which is sent back to the user. The loop depicted in Fig. 1 is repeated until the application completes the modelled task.

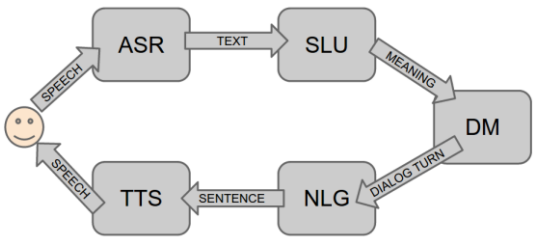


Figure 1. SDS structure.

B. How SLU Works

SDS needs a sophisticated SLU module [9] in order to implement dialog applications that go beyond solving simple tasks like call routing or form filling [11]. SLU is performed as a semantic parsing of spoken sentences. Current works in language modelling focus on two main areas: formal and stochastic approaches. Formal approaches to language modelling come in many forms and serve many motivations. This problem relates to the hand-coding of definitions of a language. Stochastic approaches involve the compilation of a finite-state machine in which the likelihood of a given word occurring is calculated based on the corpus, possibly, having the context of the preceding *n* words. All the stochastic models for SLU proposed to so far, perform the translation from a spoken sentence to a semantic constituent-based representation using statistical learning models. These systems are TINA [12], Chronus system from AT&T [13].

Development time, reusability and expertise required to create the language model, play a role in determining an appropriate solution in many cases [14]. Furthermore, manually developed grammars require combined linguistic and engineering expertise to construct a grammar with good coverage and, therefore, performance. It takes multiple rounds to fine tune a grammar, and it is difficult and expensive to maintain it. The second research paradigm adopts a data-driven, stochastic modelling approach. While it alleviates the labor-intensive problem associated with the first paradigm, it requires a huge amount of training data, which is seldom available for industrial applications.

These are difficulties [15] and the research community has potential areas of improvement focusing on these two problems:

- Systems have to be developed with little or no data. The manual grammar authoring is necessary for initial system deployment. Tools for fast grammar handcrafting make easier to enlarge the coverage of a grammar and, therefore, are crucial in this case.
- There are huge amounts of data available after deployment. It is hard to manage and manually analyze the data in order to find the problems in the initial deployment.

In this article, we introduce a grammar-authoring tool represents a solution for the first problem.

III. SLU IN NEO-LATIN LANGUAGES

All Romance languages have common features, so one could imagine a SLU system that takes into account these characteristics and shows the same behavior for such languages. This section explains some features of the Romance languages.

The Neo-Latin or Romance languages are the direct continuation of Latin, a language with a very rich dictionary. In fact, Latin has a high level of perfection as it was the idiom of a population with an advanced degree of civilization [16]. Today, many voices have disappeared;

instead, others are present in the Romance languages [17]. The Latin lexicon had been always in continuous evolution and at the time it had become wealthy of new elements, at times taken from foreign languages, but mainly through the addition of suffixes, for example to create diminutive forms. The creation of new forms through the addiction of suffices is also a characteristic of the Romance languages, in fact, based on a word that has a particular suffix, infinite other words can be formed.

Today, there are many Neo-Latin languages, the main are: Italian, Portuguese, Spanish, French, Provenzal and Romanian.

A characteristic of the Romance languages is, as in Latin, the creation of inflections. The Romance languages are highly inflectional, in which each inflection does not change the part of speech category but the grammatical function. In general, the inflected forms are obtained by adding to the root of a canonical form a particular desinence (but there are some irregular cases in which also the root changes, this phenomenon is called apophony).

Conjugations are inflections of verbs; they provide information about mood, tense, person, number (singular or plural) and gender (masculine or feminine) in the past participle. Declensions are inflections of nouns and adjectives; they provide information about gender and number.

The conjugations, which in Latin are four, in Romance languages are three. According to the conjugations, different declensions are applied: the first conjugation is for verbs that end in "-are", the second is for verbs that end in "-ere", the third is for verbs that end in "-ire" (in Latin, there is a distinction between -ĒRE and -ĔRE).

In the transition from Latin to the Romance languages, in some cases there have been passages of conjugation (called "metaplasmi"). In general, in verbs moods and temps have not changed, but there are disappeared or innovated forms for function or meaning. The disappeared forms are: deponent verbs (that are verbs with passive form and active meaning), simple future (the simple future of the Romance languages is not derived from Latin), perfect subjunctive, future imperative, future infinitive, supine, and gerundive.

The alterations have occurred for various reasons, for example, for phonetic problems, many "b" were turned into "v", because their pronunciation was very similar (e.g., "cantabit" in Italian becomes "cantavi"). In Latin verbs, many tenses have similar declensions (e.g., the future perfect and the perfect subjunctive, the subjunctive and the infinite present). As a result, many verbal forms have disappeared (e.g., "supine", gerundive, declensions of infinitive, future participle) and have been replaced by forms that are more expressive. In this way, new verb forms were born.

To form the future, different periphrastic constructions were created, for example, the most common is derived from the union of the infinitive and the reduced forms of the present indicative of "habere", with the accent on the auxiliary verb (e.g., the Latin form "cantābo" becomes "canterò" in Italian, "chanterai" in French, "cantaré" in Spanish).

The conditional does not exist in Latin and in the Romance languages it is derived from the union of infinitive and the reduced forms of perfect or imperfect of "habere" (e.g., "canterei" in Italian, "chanterais" in French, "cantaria" in Spanish).

Periphrastic forms with the past participle, as passive forms and all the compound tenses, are typical of the Romance languages (e.g., the Latin form "amor" becomes "io sono amato" in Italian, "je suis aimé" in French).

There are Latin verb forms that have transformed their function, for example the pluperfect subjunctive has the meaning of imperfect subjunctive (e.g., the Latin "cantavissem", that meant "avessi cantato" in Italian, now means "cantassi", "chantasse" in French, "cantase" in Spanish); this happened because the imperfect subjunctive in Latin ("cantarem") was too similar to the present infinitive.

Therefore, the Romance languages have a very similar way to create inflections of verb, nouns and adjective and suffixed forms. This allows creating, for these idioms, similar algorithms for generation and morphological analysis.

A. Italian Language

Like all the Romance languages, Italian is highly inflectional. Italian has three conjugations for verbs, each conjugation involves the application of specific suffixes: the verbs that end in "-are" belongs to the first conjugation, the verbs that end in "-ere" belongs to the second and the verbs that end in "-ire" to the third. Each inflected form of a verb gives information about mood, temp, number and person, and gender and number in the case of the participle. In Italian, there are many irregular verbs. Irregular verbs that end in "-rre" belong to the second conjugation. Italian irregular forms often originate from Latin irregular forms.

Latin had five declensions of nouns and adjectives, which have undergone a significant rearrangement. In Italian nouns and adjectives create inflections in various ways, for example to form the plural some nouns remain the same, others have many plural, sometimes with different meanings. Many nouns are irregular when the gender changes. Nouns and adjectives are subject to alteration that is the addition of a suffix to change the meaning in evaluation, quantity or quantity. Adverbs are not inflected and can be obtained by adding a particular suffix to some adjectives.

Italian has many orthographic rules related to its phonetic. Italian words can be reproduced by the combination of 28 different sounds called phonemes. There is not always a correspondence between phonemes and letters, in fact, some letters represent different sounds according to the following vowel [18]. For example, if "c" and "g" are followed by the vowels "a", "o" and "u", they produce a hard sound and if the vowels "e" and "i" follow them they produce a soft sound. To obtain the corresponding hard sound the letter "h" is inserted between these characters and vowels "e" and "i"; to obtain the soft sound with the vowels "a", "o" and "u" the character "i" is inserted.

There are other orthographic rules that concern the behavior of groups of two or three characters as "sc", "gn" and "gl".

IV. OLYMPUS

This work was designed and tested within the framework Olympus [19]. Olympus is a complete framework for implementing SDSs created at Carnegie Mellon University (CMU) during the late 2000's. Olympus includes a dialog manager called RavenClaw [20], which supports mixed-initiative interaction, as well as NLU components that handle speech recognition (Sphinx) and understanding (Phoenix). Olympus uses a Galaxy [21] message-passing layer architecture to integrate its components and supports multi-modal interaction. The Galaxy architecture is a set of Galaxy *Servers*, which communicate to each other through a central Galaxy module called *Hub*. Olympus provides the infrastructure upon which it is possible to build Spoken Dialog Applications. Specific application functions such as instance dialog planning, input processing, output processing and error handling are encapsulated in subcomponents with well-defined interfaces that are decoupled from domain-specific dialog control logic. Each application needs the following domain specific components: a specific grammar, a dialog manager, a back-end server and a language generator module. These modules are strictly domain dependent and represent the core of the spoken interaction. The Phoenix parser represents the NLU module in the Olympus framework. The Phoenix parser [8] was developed by the University of Colorado in 2002 to develop easy and robust Natural Language Processing systems. It was then adopted by the CMU and used in the Olympus framework. The parser performs the human language syntactic analysis according to the rules that are defined in its grammar. For each user input sentence, the Sphinx module of the Olympus framework produces *n* text output. Each of them is associated with a probability. The higher is the probability, the more likely is the association between a text and a user sentence. Each of these *n* texts is parsed by Phoenix. The meaning extracted from the input sentence will then direct the Dialog Manager in deciding the corresponding action. Subsequently, the Natural Language Generation module will produce the output sentence.

A. SLU grammar in Olympus Framework

To build a specific Spoken Language Application in the Olympus Framework the developer has to design and develop specific grammar definition. The Phoenix parser uses a formal method and a hand crafted CFG Grammar. It requires combined linguistic and engineering expertise to construct a grammar with good coverage and optimized performance. First of all, the developer has to determine the main set of jobs that the application will handle. Each concept or action defined in the dialog manager is mapped in one or more grammar slots. Therefore, the design of the grammar is strictly bound to the design of the dialog tree. Grammar rules are specified in the source grammar file. The manual development of Phoenix grammars is a time-consuming and tedious process that requires human expertise, posing an obstacle to the rapid porting of SDS to new domains and languages. A semantically coherent workflow for SDS grammar development starts from the definition of low-level rules and proceeds to high-level ones. The Olympus framework provides English generic grammar files, which contains some standard forms such as greetings, social expressions and yes/no, as well as discourse entities such as help, repeat, etc. This grammar has to be extended by introducing domain-specific phrases.

B. Phoenix Grammar

Since spontaneous speech is often ill formed and the recognizer makes errors, it is necessary that the parser is robust to recognition errors, grammar and fluency. This parser is designed to enable robust parsing of these types of input. The Phoenix parser uses a specific CFG grammar that is organized in a grammar file. Names of grammar files end with a ".gra" extension. This contains context-free rules that specify the word patterns corresponding to the token. The syntax for a grammar for a token is in Fig. 2. In Fig. 3 there is an example.

```
# optional comment
[token_name]
      (<pattern a>)
      (<pattern b>)
;
```

Figure 2. Generic Phoenix grammar syntax.

```
[city]
      (New York)
      (London)
      (Paris)
;
```

Figure 3. Example of a Phoenix token.

A token can also contain other tokens, for example (Fig. 4):

```
[token_example]
      (word1 [other_token] word2)
;
```

Figure 4. Example of a Phoenix token containing other tokens.

This format allows recognizing several sentences with the combination of different slots and words; furthermore, each token can be reused in many tokens.

In the inflective languages, as Italian or Romance languages in general, words can occur in several forms, verbs can change its form depending on conjugations and nouns and adjectives depending on declensions. Their forms can change also applying different suffixes or prefixes. This means that the Phoenix grammar must contain all the possible inflected forms. For this reason, the grammar can become long and hard to write, because the developer must manually write it and he might forget some inflected forms: the result can be a not complete grammar. This increases the development time.

Thus, inflected forms add complexity to the Phoenix grammar, since they generate multiple different rules with similar patterns.

V. A NEW GRAMMAR FORMAT

The development of a new domain application needs a new Context Free Grammar (CFG) that is able to define the concepts and their relations of such domain.

Alternative approaches learn structures from a set of corpora. However, this process appears too expensive and potentially not exhaustive [22].

Our approach consists of creating a new intermediate grammar that focuses on the meaning of a grammar token rather than on its content.

The legal combination of individual words into constituents and constituents into sentences represents a semantic context free grammar (CFG).

When a regular software developer develops a new application for a new domain, he must define a new grammar by a CFG that is able to define the concepts and their relations of such domain.

Even if other approaches suggest learning structures from a set of corpora, this process appears too expensive and probably not exhaustive [23], our solution can facilitate grammar development by supporting the flow of information from a manually written source to language contents automatically generated.

The goal is to make sure that the programmer needs only to think about the meaning of a grammar token and not about their content. This new schema, thanks to a Morphological Generator [24], generates a file that can be reused and edited like a standard phoenix grammar.

The new format description is in Fig. 5.

```
Function = NAME_ACTION
{
      [token1] term1 [token2] term2
}
;
```

Figure 5. New grammar syntax.

A set of slots, that represent information that is relevant to the action or object (in this case "NAME_ACTION"), are defined by the "Function" keyword that defines the tag name (Function = NAME_ACTION).

The content between curly brackets is described by a new grammar tag definition mode. The number and the order of tokens and terms can change. Each token is written in square brackets.

In such a way, it is no longer necessary to write the word pattern of the token, but only the "keyword name" like [word, characteristic].

The new schema creates a grammar file containing a token and its generated word patterns and it can be reused and edited like a standard Phoenix grammar. Fig. 6 depicts the new format description.

```
Function = SLOT_NAME
{
      [word,characteristic] term1
      [word,characteristic] term2
}
;
```

Figure 6. New grammar format description.

The grammar slots are defined by the "Function" keyword that defines the slot name (Function = SLOT_NAME). In this way, a tag is defined as a couple "[word, characteristic]" and it is used by the editor to generate the appropriate **word** patterns according to the **characteristic**.

The couple [word, characteristic] is defined as below:

- If "word" is a verb, "characteristic" can be replaced with:
  - "presente" if the Italian present forms are desired;
  - "passato" if the Italian past forms are desired;
  - "futuro" if the Italian future forms are desired.
- If "word" is a noun or an adjective, "characteristic" can be replaced with:
  - "singolare" if the Italian singular forms are desired;
  - "plurale" if the Italian plural forms are desired;

Our version of the Morphological Generator generates all new forms specified by the characteristic.

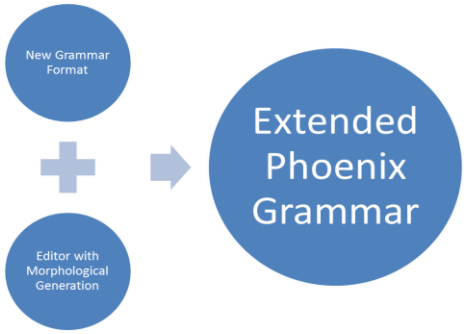


Figure 7. New grammar generation.

Our editor generates an extended standard Phoenix grammar with increased coverage of the new grammar, by performing the following actions:

- Creation of a token named SLOT_NAME in which new tokens and terms are included;
- Creation of a token for each new defined token, in which terms generated by the Morphological Generator are included.

Fig. 7 shows the full process.

Our tool consists of the Editor component, which takes the new grammar as an input and, with the aid of the

Morphological Generator, generates the grammar format for Phoenix. The following paragraphs explain in detail the other components.

VI. MORPHOLOGICAL GENERATOR FOR THE ITALIAN LANGUAGE

The Morphological Generator allows you to generate specific inflected and altered forms of nouns, adjectives and verbs. It is a fundamental tool as it allows generating the inflected forms of the language supported.

Since each lemma follows different rules for the creation of the inflections, the Morphological Generator uses a word-list in which a *grammatical category* is associated to each lemma, according to the following format:

```
lemma,grammatical_category;
```

The grammatical category is a string that contains information about the part of speech of the lemma and its way of creating inflections.

For the verbs, there are four grammatical categories:

- one for the intransitive verbs (VI);
- one for transitive verbs (VT);
- one for auxiliary verbs (VA);
- one for modal verbs(VS).

Suffixes for the different conjugation are chosen by analyzing the last three characters with which the verbal lemma ends: these determine the *verbal group code*. In this way, if the verbal lemma ends in "-are", the suffixes of the first conjugation are applied; if it ends in "-ere" or "-rre", the suffixes of the second conjugation are applied; if it ends in "-ire" suffixes of the third conjugation are applied. If the verb is irregular, the grammatical category contains also an *inflectional code* that is a number that allows deriving irregular inflections.

There are many grammatical categories of nouns and adjectives. For example, there is a grammatical category of neuter nouns that can generate four different inflectional forms (one of the masculine singular, one of the feminine singular, one of the masculine plural, one of the feminine plural), another of feminine nouns, another of masculine forms, another of neuter nouns that have invariable feminine forms, and so on; similarly for the adjectives. Inflections are chosen because of the grammatical category and the last characters with which the lemma ends, which determines the *noun group code* (for nouns) or the *adjectival group code* (for adjectives). In fact, to each grammatical category of nouns and adjectives some rules are associated.

There is also a grammatical category of irregular nouns and one for irregular adjectives; these do not follow rules to create the inflections, so the inflections are obtained from a specific list that contains all irregular forms.

For the other parts of speech, there are the following grammatical categories:

- E for prepositions;
- C for conjunctions;
- B for adverbs;
- R for articles;
- P for pronouns;

For these lemmas, inflections are not applied.

Fig. 8 shows the algorithm; the lemma is the input and the list of all obtained inflected forms is the output. If the lemma is not declinable, the output is the same lemma in input.

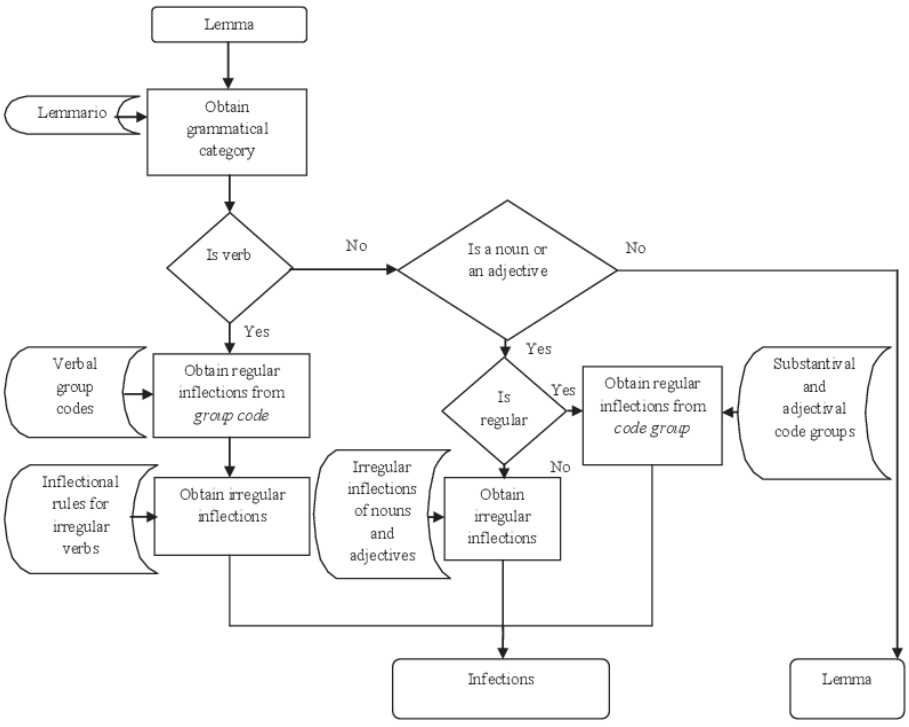


Figure 8. Morphological Generator.

In Italian, nouns and adjectives can be altered by adding particular suffixes. The alteration modifies the meaning of a word in quantity or quality. The Morphological Generator applies 9 adjectival suffixes for the alteration, each of which can be inflected in gender and number, so in total there are 36 (9x4) possible altered adjectives. There are also 8 substantival suffixes for the alteration, each of which can be inflected, so in total there are 32 (8x2) possible altered nouns. Furthermore, 7 prefixes can be applied to all forms of nouns and adjectives.

When inflectional suffixes are applied, orthographic rules for Italian are respected. The Italian words can be uttered by the combination of many sounds, but sometimes there are not correspondence between the sound and the characters, in fact, some letters have different sounds according to the vowel that follows them. When the suffix of the lemma is removed, the root is obtained and in general, the following rules are applied:

- if the root ends in “-c” or “-g”:
  - if the desinence of the canonical form is “-a”, “-o” or “-u” (forming with the root an hard sound) and the suffix to be applied starts in “e” or “i”, the character “h” is inserted before the suffix.
  - if the desinence of the canonical form is “-e” or “-i” (forming with the root a soft sound) and the suffix to be applied starts in “a”, “o” or “u”, the character “i” is inserted before the suffix.
- the vowel “i” is removed from the root if:
  - the root ends in “-ci” or “-gi” and the suffix starts in “e”;
  - the root ends in “-i” and the suffix starts in “i”.

There are also particular words that not follow these rules. In these cases, the words belong to a particular grammatical category that nullifies the rules above. Furthermore, there are particular orthographic rules for verbs.

Each generated word is stored in a structure that saves information about the inflection: part of speech, mood, temp, gender, number, suffix applied and prefix applied. Therefore, the algorithm is able to give in output only the inflections of a lemma required by the user (for example, all the past tenses of a verb or the singular forms of a noun or an adjective). This characteristic is used for the generation of the new grammar.

This method can apply not only to the Romance languages. It can be applied to others inflective languages. For example, many Morphological Generators [26][27][28], one for a given language, can be utilized and the editor can generate many Phoenix grammar files, one for each language. In this way, the developer writes the grammar in a single language and obtains a multilingual result.

VII. GRAMMAR EDITOR

The grammar editor (Fig. 9) consists of a text editor modified for our purposes. This editor supports the new grammar format and the user produces the corresponding .gra file, by clicking on the "generate" button.

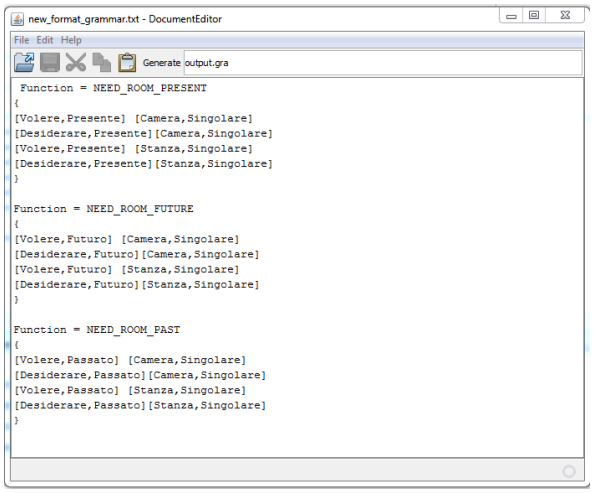


Figure 9. GUI of the editor.

This component reads and processes the grammar files (new format) and, using the Morphological Generator, obtains a Phoenix grammar file (Fig. 10).

Figure 10. New grammar generation process.

If the programmer does not know the SDS domain language, he enables the "translator" module (Fig. 11) between the Morphological Generator and the Editor.

For example, an English language grammar, as shown in Fig. 12, is translated by a component of the Editor into the target language and then is used by the Morphological Generator to generate the grammar of the target language in the Phoenix grammar format.

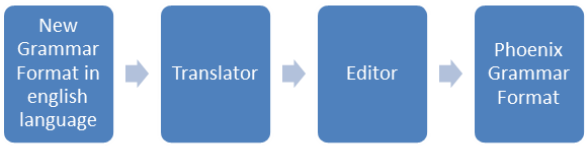


Figure 11. New grammar generation process with translator.

In this way, we have a grammar written in a universal language (English) with a high level of abstraction that can generate more coverage of a grammar written by a programmer in the same time. In addition, Phoenix grammars in different languages that are not initially known by the regular software developer.

Each grammar that is produced requires a different morphological generator.

```
Function = NEED_ROOM_PRESENT
{
      [Will,Present] [Room, Singular]
      [Want,Present] [Room, Singular]
      [Will,Present] [Room, Singular]
      [Want,Present] [Room, Singular]
}
```

Figure 12. New grammar written in English language.

The complexity of the grammar of Italian and Neo-Latin languages, in general, increases the effort in developing an efficient SLU grammar for a SDS.

With this system, the regular software developer can generate a Phoenix grammar without worrying about all the possible variations, conjugations and alterations of words that are characteristics of the Romance languages

VIII. Experimental Results

An example is reported to show the advantages brought by this approach. It shows a Phoenix grammar of a real SDS for a room-reservation application, based on the Olympus Framework. In a typical interaction, the user can express the same concept using a specific word, but in different tenses. For example, “I want a room” in Italian can be expressed like “Voglio una camera”, but also “Vorrei una camera” (“I'd like to have a room”) or “Vorrei una cameretta” (“I'd like to have a small room”, in Italian it is a term of endearment). Fig. 11 shows an example of grammar.

```
Function=NEED_ROOM_PRESENT
{
      [volere,presente][camera,singolare]
      [desiderare,presente][camera,singolare]
      [volere,presente][stanza,singolare]
      [desiderare,presente][stanza,singolare]
}

Function=NEED_ROOM_FUTURE
{
      [volere,futuro][camera,singolare]
      [desiderare,futuro][camera,singolare]
      [volere,futuro][stanza,singolare]
      [desiderare,futuro][stanza,singolare]
}
```

Figure 13. Example of Italian grammar.

The new grammar consists of two parts. The first one, shown in Fig. 12, represents the definition of a grammar slot.

```
[NEED_ROOM_PRESENT]
      [volere_presente][camera_singolare]
      [desiderare_presente][camera_singolare]
      [volere_presente][stanza_singolare]
      [desiderare_presente][stanza_singolare]
;

[NEED_ROOM_FUTURE]
      [volere_futuro][camera_singolare]
      [desiderare_futuro][camera_singolare]
      [volere_futuro][stanza_singolare]
      [desiderare_futuro][stanza_singolare]
;
```

Figure 14. Generated grammar slots.

The second part, shown in Fig. 13, defines each token, including their word patterns. A more detailed explanation is along with the source code (output.gra file) [25].

The initial grammar, consisting of 21 rows, generates a 140-row-long Phoenix grammar that allows the SLU module to recognize a large set of utterances.

This way, the developer focuses his attention on the meaning of an intermediate-grammar token and not on its content.

```
#tag auto generated
[volere_presente]
      (voglio)
      (vuoi)
      ...
;

#tag auto generated
[volere_futuro]
      (vorrò)
      (vorrai)
      ...
;

#tag auto generated
[stanza_singolare]
      (stanza)
      (stanzaccia)
      ...
;

#tag auto generated
[camera_singolare]
      (camera)
      (cameraccia)
      ...
;

#tag auto generated
[desiderare_presente]
      (desidero)
      (desideri)
      ...
;

#tag auto generated
[desiderare_futuro]
      (desidererò)
      (desidererai)
      ...
;
```

Figure 15. Generated tokens.

Furthermore, the developer does not need to write all possible forms (mood, tense, person, etc.), some of which could be difficult to predict. The advantage of the generated grammar is the ability to easily simulate and predict the large variety of interactions that can occur.

The same grammar can also be obtained starting from an initial grammar written in another language, for example, in English, and enabling the translator module, as shown in Fig. 14.

```
Function=NEED_ROOM_PRESENT
{
      [to want,present][room,singular]
      [to desire,present][room,singular]
      [to want,present][apartment,singular]
      [to desire,present][apartment,singular]
}

Function=NEED_ROOM_FUTURE
{
      [to want,future][room,singular]
      [to desire,future][room,singular]
      [to want,future][apartment,singular]
      [to desire,future][apartment,singular]
}
```

Figure 16. Example of English grammar.

The generated grammar slots are shown in Fig. 15 and the associated tokens in Fig. 16.

```
[NEED_ROOM_PRESENT]
      [to_want_present][room_singular]
      [to_desire_present][room_singular]
      [to_want_present][apartment_singular]
      [to_desire_present][apartment_singular]
;

[NEED_ROOM_FUTURE]
      [to_want_future][room_singular]
      [to_desire_future][room_singular]
      [to_want_future][apartment_singular]
      [to_desire_future][apartment_singular]
;
```

Figure 17. Generated grammar slots from English.

```
#tag auto generated
[to_want_present]
      (voglio)
      (vuoi)
      ...
;

#tag auto generated
[to_want_future]
      (vorrò)
      (vorrai)
      ...
;

#tag auto generated
[room_singular]
      (stanza)
      (stanzaccia)
      ...
;

#tag auto generated
[apartment_singular]
      (camera)
      (cameraccia)
      ...
;

#tag auto generated
[desire_present]
      (desidero)
      (desideri)
      ...
;

#tag auto generated
[desire_future]
      (desidererò)
      (desidererai)
      ...
;
```

Figure 18. Generated tokens from English.

IX. Conclusion and Future Work

This paper investigates the problem of grammar authoring for initial system deployment when little data is available.

In this work, we propose a solution to simplify and reduce the amount of writing of the SDS grammar of inflectional language. This method reduces the effort to produce a grammar for a SDS especially for a regular software developer. The SDS used for our tests is the Olympus framework.

An editor has been developed for the translation of the new simple grammar format in the Phoenix grammar format. The editor uses a new Morphological Generator to obtain all possible inflected words that are used to create grammar tokens.

The proposed solution will be integrated in a major project called Olympus P2P [29], which is concerned with the upgrading and updating of an SDS grammar by means a peer-to-peer network to share new grammar tokens generated from the new grammar format.

ACKNOWLEDGMENT

The authors were supported by the Sicilian Region grant PROGETTO POR 4.1.1.1: "Rammar Sistema Cibernetico programmabile d'interfacce a interazione verbale".

REFERENCES

[1] S. M. Biondi, V. Catania, Y. Cilano, R. Di Natale and A.R. Intilisano, "An Easy and Efficient Grammar Generator for Spoken Language Understanding," The Sixth International Conference on Creative Content Technologies (CONTENT) pp 13-16, Venice, May 2014.

[2] I. Klasinas, A. Potamianos, E. Iosif, S. Georgiladakis, and G. Mameli, "Web data harvesting for speech understanding grammar induction," in Proc. Interspeech, Lyon, France, Aug. 2013.

[3] J. F. Allen, B. W. Miller, E. K. Ringger, T. Sikorshi, "Robust understanding in a dialogue system," 34th Annual Meeting of the Association for Computational Linguistics. Santa Cruz, California, USA, pp. 62–70, 1996.

[4] S. Bangalore, M. Johnston, "Balancing data-driven and rule-based approaches in the context of a multimodal conversational system," Human Language Technology/Conference of the North American Chapter of the Association for Computational Linguistics. Boston, MA, USA, 2004.

[5] H. M. Meng and K-C. Siu, "SemtiautomaticAcquition of Semantic Structures for Understanding Domain-Specific Natural Language Queries," IEEE Tran. Knowledge & Data Eng., pp. 172–181, vol. 14(1), 2002.

[6] Y. Wang and A. Acero, "Grammar learning for spoken language understanding," Automatic Speech Recognition and Understanding, ASRU '01. IEEE Workshop, pp. 292-295, 2001.

[7] Y.-Y. Wang and A. Acero, "Rapid development of spoken language understanding grammars," Speech Communication, vol. 48, no. 3-4, p. 390416, 2008.

[8] W. Ward, "Understanding spontaneous speech: the Phoenix system," Acoustics, Speech, and Signal Processing, ICASSP-91, 1991 International Conference, pp. 365-367 vol. 1, 14-17 Apr 1991.

[9] Phoenix Parser User Manual, http://www.ontolinux.com/community/phoenix/Phoenix_Manual.pdf (last visited: 22 November 2013).

[10] Phd Thesis, International Doctorate School in Information and Communication Technologies DISI - From Spoken Utterances to Semantic Structures Marco Dinarelli.

[11] A. L. Gorin, G. Riccardi, and J. H. Wright, "How may i help you?," Speech Commun., 23(1-2): 113–127, 1997.

[12] S. Seneff. Tina, "A natural language system for spoken language applications," Comput. Linguist., 18(1): 61–86, 1992.

[13] N. Cancedda, E. Gaussier, C. Goutte, and J. M. Renders, "Word sequence kernels," J. Mach. Learn. Res., 3, 2003.

[14] Language Modelling for Spoken Dialogue Systems; Grammar-Based and Robust Approaches Compared and Contrasted Genevieve Gorrell, December 22, 2003

[15] R. Pieraccini, "Spoken language understanding, the research/industry chasm," HLT-NAACL 2004 Workshop: Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Information for Speech Processing, pp 47-47, Boston, May 2004

[16] Lingue Neolatine in Treccani.it - Enciclopedie on line. Istituto dell'Enciclopedia Italiana.

[17] Grammatica Storica in Treccani.it - Enciclopedie on line. Istituto dell'Enciclopedia Italiana.

[18] F. Musso, N. Prandi, "Per dirla giusta. Fonologia, ortografia, morfologia", S. Lattes & C. Editori SpA, 2012.

[19] D. Bohus, A. Raux, T. K. Harris, M. Eskenazi, and A. I. Rudnicky, "Olympus: an open-source framework for conversational spoken language interface research," NAACL-HLT '07: Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies, 2007

[20] D. Bohus, A. I. Rudnicky, "The RavenClaw dialog management framework: Architecture and systems," Computer Speech and Language, vol. 23, no. 3, 2009

[21] J. Polifroni, and S. Seneff, "Galaxy-II as an Architecture for Spoken Dialogue Evaluation," Proc. LREC, 725-730, Athens, 2000.

[22] S. Knight, G. Gorrell, M. Rayner, D. Milward, R. Koeling and I. Lewin, "Comparing grammar-based and robust approaches to speech understanding: a case study," EUROSPEECH 2001 Scandinavia.

[23] M. Haspelmath and A. D. Sims, Understanding Morphology 2nd edition. London: Hodder Education, 2010.

[24] V. Catania, Y. Cilano, R. Di Natale, V. Mirabella and D. Panno, "A morphological engine for Italian language", ICIEET 2013: 2nd International Conference on Internet, E-Learning & Education Technologies, 2013, pp. 36-43, vol. 12(1).

[25] Source code, http://opensource.diit.unict.it/vctsds/GrammarEditor.zip (last visited: 22 November 2013).

[26] Anandan, P., Ranjani Parthasarathy & Geetha, T.V., "Morphological Generator for Tamil," Tamil Internet Conference, Kuala Lumpur, Malaysia, 2001.

[27] George Petasis, Vangelis Karkaletsis, Dimitra Farmakiotou, George Samaritakis, Ion Androutsopoulos, Constantine D. Spyropoulos , "A Greek Morphological Lexicon And Its Exploitation By A Greek Controlled Language Checker," In Proceedings of the 8th Panhellenic Conference on Informatics, pp. 8 – 10, 2001.

[28] Habash, N., Ower, R., and George, "Morphological analysis and generation for Arabic dialects," Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, pages 17–24, Ann Arbor, June 2005.

[29] V. Catania, R. Di Natale, A. Longo and A. Intilisano, "A distributed Multi-Session Dialog Manager with a Dynamic Grammar Parser," 2nd International Conference on Human Computer Interaction & Learning Technologies, 2013, pp. 1-9, vol. 8(2).

Intelligent Search Engine to a Semantic Knowledge Retrieval in the Digital Repositories

Antonio Martín
Department of Electronic Technology
Higher Technical School of Computer Engineering
Sevilla, Spain
toni@us.es

Abstract— **Currently, an enormous quantity of heterogeneous and distributed information is stored in the current digital libraries. This data abundance has made the task of locating relevant knowledge more complex. Such complexity drives the need for intelligent systems for searching and for knowledge retrieval. Access to these collections poses a serious challenge. The present search techniques based on manually annotated metadata and linear replay of material selected by the user do not scale effectively or efficiently to large collections. The Artificial Intelligence and Semantic Web provide a common framework that allows knowledge to be shared and reused. In this paper, we propose a comprehensive approach for discovering information objects in large digital collections. The process is based on analysis of recorded semantic metadata in those objects and the application of expert system technologies. We suggest a conceptual architecture for a semantic and intelligent search engine. We concentrate on the critical issue of metadata/ontology-based search. More specifically, the objective is investigated from a search perspective possible intelligent infrastructures form constructing decentralized digital libraries where no global schema exists. We have used Case Based-Reasoning methodology to develop a prototype for supporting efficient retrieval knowledge from digital library of Seville University. The work suggests a conceptual architecture for a semantic and intelligent search engine and we also have developed a prototype and tested it for supporting efficient retrieval knowledge from digital libraries.**

Keywords-Ontology; Semantic Web; Retrieval; Case-based Reasoning; Digital Library; Knowledge Management.

I. INTRODUCTION

A Digital Library (DL) enables users to interact effectively with information distributed across a network. These network information systems support search and display of items from organized collections. In the historical evolution of digital libraries the mechanisms for retrieval of scientific literature have been particularly important. Traditional search engines treated the information as an ordinary database that manages the contents and positions. The result generated by the current search engines is a list of Web addresses that contain or treat the pattern. The useful information buried under the useless information cannot be discovered. It is disconcerting for the end user. Thus, sometimes it takes a long time to search for needed information. Although search engines have developed increasingly effective, information overload obstructs precise searches. Despite large investments and efforts have been made, there are still a lot of unsolved problems. Thus, it is necessary to develop new intelligent and semantic models that offer more possibilities [1].

There are researchers and works in related fields, which include ontology retrieval methods. The study [2] presents a system, which uses an ontology query model to analyze the usefulness of ontologies in effectively performing document searches. This work proposes an algorithm to refine ontologies for information retrieval tasks with preliminary positive results. [3] uses a medical ontology to improve a Multimodal Information Retrieval System by expanding the user's query with medical terms. The study [4] combines swarm intelligence and Web Services to transform a conventional library system into an intelligent library system with high integrity, usability, correctness, and reliability software for readers. The research [5] proposes meta-concepts with which the ontology developers describe the domain concepts of parts libraries. The meta-concepts have explicit ontological semantics, so that they help to identify domain concepts consistently and structure them systematically. The study [6] presents a formulation and case studies of the conditions for patenting content-based retrieval processes in digital libraries, especially in image libraries. The paper [7] focuses on methods for evaluating different symbolic music matching strategies, and describes a series of experiments that compare and contrast results obtained using three dominant paradigms. The research [8] proposes organizational memory architecture and annotation and retrieval information strategies. This technique is based on domain ontologies that take in account complex words to retrieve information through natural language queries.

There are a lot of researches on applying these new technologies into current information retrieval systems, but no research addresses Artificial Intelligence (AI) and semantic issues from the whole life cycle and architecture point of view [9]. Although search engines have developed increasingly effective, information overload obstructs precise searches. Our work differs from related projects in that we build ontology-based contextual profiles and we introduce an approaches used metadata-based in ontology search and expert system technologies [10]. We presented an intelligent approach for optimize a search engine in a specific domain. This study improves the efficiency methods to search a distributed data space like DL. The objective has focused on creating technologically complex environments digital repositories domain. It incorporates Semantic Web and AI technologies to enable not only precise location of public resources but also the automatic or semi-automatic learning [11].

Our approach for realizing content-based search and retrieval information implies the application of the Case-Based Reasoning (CBR) technology [12]. Thus, our objective here is to contribute to a better knowledge retrieval in DL field. This paper describes semantic interoperability problems and presents an intelligent architecture to address them, called OntoSDL. Obviously, our system is a prototype but, nevertheless, it gives a good picture of the on-going activities in this new and important field. We concentrate on the critical issue of metadata/ontology-based search and expert system technologies. More specifically, the objective is investigated from a search perspective possible intelligent infrastructures form constructing decentralized public repositories where no global schema exists.

The contributions are divided into next sections. In the first section, short descriptions of important aspects in DL domain, the research problems and current work in it are reported. Then, we summarize its main components and describe how can interact AI and Semantic Web to improve the search engine. Third section focuses on the ontology design process and provides a general overview about our prototype architecture. Next, we study the CBR framework jColibri and its features for implementing the reasoning process over ontologies [13]. Finally, we present conclusions of our ongoing work on the adaptation of the framework and we outline future works.

II. MOTIVATIION AND REQUIREMENTS

In the historical evolution of DL, the mechanisms for retrieval information and knowledge have been particularly important. These network information systems support search and display of items from organized collections. Reuse this knowledge is an important area in this domain. The Semantic Web provides a common framework that allows knowledge to be shared and reused across community users [14].

Repositories and digital archives are privileged area for the application of innovative, knowledge intensive services that provide a flexible and efficient method for searching information and guarantee the user with a set of results actually related to his/her interest. Seville University institutional repository is dedicated to the production, maintenance, delivery, and preservation of a wide range of high-quality networked resources for citizens, scholars, and students at University and elsewhere. This repository includes services to effectively share their materials and provide greater access to digital content [15].

Thus, the goal is to contribute to a better knowledge retrieval in the institutional repositories dominium. This scheme is based on the next principles: knowledge items are abstracted to a characterization by metadata description, witch is used for further processing. This characterization is based on a vocabulary/ontology that is shared to case the access to the relevant information sources. This begets new challenges to docent community and motivates researchers to look for intelligent information retrieval approach and ontologies that search and/or filter information automatically based on some higher level of understanding are required. We make an effort in this direction by investigating techniques that attempt to utilize ontologies to improve effectiveness in information retrieval. Thus, ontologies are seen as key enablers for the Semantic Web. We have proposed a method to efficiently search for the target information on a digital repository network with multiple independent information sources [16]. The use of AI and ontologies as a knowledge representation formalism offers many advantages in information retrieval [17]. In our work, we analysed the relationship between both factors ontologies and expert systems.

We focus our discussion on case indexing and retrieval strategies and provide a perception of the technical aspects of the application. For this reason, we are improving representation by incorporating more metadata within the information representation [18]. We discuss an opportunity and challenge in this domain with a specific view of intelligent information processing that takes into account the semantics of the knowledge items. In this paper, we study architecture of the search layer in this particular dominium, a web-based catalogue for the University of Seville. The hypothesis is that with a case-based reasoning expert system and by incorporating limited semantic knowledge, it is possible to improve the effectiveness of an information retrieval system [19]. More specifically, the objectives are decomposed into:

- Explore and understand the requirements for rendering semantic search in an institutional repository.
- Investigate how semantic technologies can be used to provide additional semantic properties from existing resources.
- Analyse the implementation results and evaluate the viability of our approaches in enabling search in intelligent-based digital repositories.

To reach these goals we need to consider information interoperability. In other words, the capacity of different information systems, applications and services to communicate, share and interchange data, information and knowledge in an effective and precise way. As well, in order to deliver new electronic products and services, ontologies can be used to integrate with other systems, applications and services. DL initiatives, such as interoperability between public services, require establishing collaborative semantic repositories among public and private sector organizations. Particularly, we require Semantic Interoperability, which is one of the key elements of the programme to support the set-up of the European E-Government services.

III. INTEROPERABILITY REQUIREMENTS

In June 2002, European heads of state adopted the Europe Action Plan 2005 at the Seville summit. It calls on the European Commission to issue an agreed interoperability framework to support the delivery of European E-Government services to citizens and

enterprises. This recommends technical policies and specifications for joining up public administration information systems across the EU. This research is based on open standards and the use of open source software. These aspects are the pillars to support the European delivery of E-Government services of the recently adopted European Interoperability Framework (EIF) [20] and its Spanish equivalent [21]. This document is reference for interoperability of the new Interoperable Delivery of Pan-European E-Government Services to Public Administrations, Business and Citizens programme (IDAbc). European Institutions and agencies should use the European interoperability framework for their operations with each other and with citizens, enterprises and administrations in the respective EU Member States [22]. Member States Administrations must use the guidance provided by the EIF to supplement their national E-Government Interoperability Frameworks with a pan-European dimension and thus enable pan-European interoperability

In this context, interoperability is the ability of information and communication technology systems and of the business processes they support to exchange data and to enable sharing of information and knowledge. The ISO/IEC 2382 Information Technology Vocabulary defines interoperability as the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units. An interoperability framework can be described as a set of standards and guidelines, which describe the way in which organisations have agreed, or should agree, to interact with each other.

Interoperability can be considered on very different abstraction levels, and the distinctions to be made in this respect cut across all the other matrix dimensions. Within a continuum ranging from a very concrete to a very abstract perspective it is possible to distinguish three layers as shown in next Fig. 1.

The aspects of interoperability as a general concept or approach cover technical, semantic, and organisational issues, usually referenced as interoperability layers. Interoperability is conceived on different main abstraction levels:

1) Organisational interoperability level: processes, defined as workflow sequences of tasks, integrated in a service-oriented environment.

2) Technical interoperability level: signals, low-level services and data transfer protocols.

3) Semantic interoperability level: information in various shared knowledge representation structures such as taxonomies, ontologies, or topic maps. Semantic interoperability is not just with about the packaging of data (data format), but mostly focuses into simultaneous transmission of their meaning (semantics). The meaning of the data is transmitted with the data itself, in an "information package" independent of any information system. Semantic interoperability shared vocabulary, and its associated links to an ontology, which provides the basis for machine interpretation and understanding of the logic of the message. This is success by adding metadata (information used to describes other data) and linking each data element to a shared vocabulary.

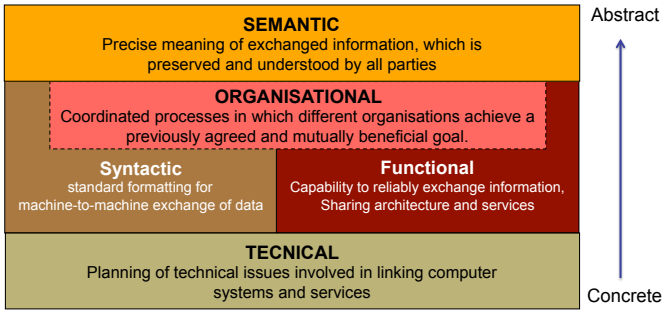


Figure 1. Abstraction layers interoperability

Two or more entities achieve interoperability when they are capable of communicating and exchanging data, which concerns to specified data formats and communication protocols. Exchanging normalized data is a prerequisite for semantic interoperability and refers to the packaging and transmission mechanisms for data. In the semantic interoperability there are concepts and methods available, but which are not yet standardized. However, for organizational interoperability it is by far less obvious what has to be standardized, who could develop and establish appropriate standards, and what is necessary for their operation and maintenance.

In this section, we have focused our work in semantic interoperability analysis. For this purpose, we use ontologies and semantic approach.

This area implies the collaboration of many actors, such as local repositories, information workers and suppliers. For this reason, we can quote the following reasons for the need to develop/define a central ontology:

- Providing a semantic typing for the data distributed all over the repositories in order to facilitate the information request by citizens through efficient search engines. Entities can be assumed to be the institutions offering digital services, digital repositories, public platforms or simply Web services.
- Sharing common understanding of the structure of information among intelligent agents, facilitating the extraction of information and processing of documents. Objects of interaction, the entities that actually need to be processed in semantic interoperability scenarios. Choices range from the full content of digital information objects to mere representations of such objects, which in turn are often conceived as metadata attribute sets.

- Enabling reuse of existing domain knowledge and its further extension, providing a contextual framework enabling unambiguous communication of complex and detailed concepts.

However, semantic interoperability problems emerge as these organizations may differ in the terms and meanings they use to communicate, express their needs and describe resources they make available to each other. Moreover, interoperability can be considered on different abstraction levels, and the distinctions to be made in this respect cut across all the other matrix dimensions. Within a continuum ranging from a very concrete to a very abstract perspective it is possible to distinguish the four layers of technical, syntactic, functional and semantic interoperability. We must bear in mind that interoperability framework is, therefore, not a static document and may have to be adapted over time as technologies, standards and administrative requirements change. In the next sections, we establish the base of all these aspects in our platform OntoSDL.

IV. THE ONTOSDL ARCHITECTURE

In order to support semantic retrieval knowledge in Seville institutional repositories we develop a prototype named OntoSDL based on ontologies and expert system technologies. The proposed architecture is based on our approach to information retrieval in an efficient way by means of metadata characterizations and domain ontology inclusion. It implies to use ontology as vocabulary to define complex, multi-relational case structures to support the CBR processes. Our system works comparing objects that can be retrieved across heterogeneous repositories and capturing a semantic view of the world independent of data representation. The framework presented in the next sections is built on established and widely accepted standards for data transfer and exchange (XML), web services (WSDL, SA-WSDL) and process models (BPMN, BPEL). The main focus of this paper is on semantic interoperability; however, other levels are addressed as well. Use of technological standards enables different kinds of interoperability constitute a major dimension with more traditional approaches geared towards librarian metadata interoperability such as Z39.50 /SRU+SRW or the harvesting methods based on OAI-PMH or again web service based approaches (SOAP/UDDI) and the Java based API defined in JCR (JSR 170/283) as well as GRID based platforms such as iRods.

The architecture of our system is shown in Fig. 2, which mainly includes three parts: ontology knowledge base, the search engine, and the intelligent user interface. Their corresponding characteristics and functions are studied in the following paragraphs.

A. Ontology Knowledge Base

OntoSDL system uses its internal knowledge bases and inference mechanisms to process information about the electronic resources in Seville University repositories. At this stage, we consider to use ontology as vocabulary for defining the case structure like attribute-value pairs. Ontology knowledge base is the kernel part for semantic retrieval information. Ontology is a knowledge structure, which identify the concepts, property of concept, resources, and relationships among them to enable share and reuse of knowledge that are needed to acquire knowledge in a specific search domain. The metadata descriptions of the resources and repository objects (cases) are abstracted from the details of their physical representation and are stored in the Case Base. Ontology provides information about resources and services where concepts are types, or classes, individuals are allowed values, or objects and relations are the attributes describing the objects [23].

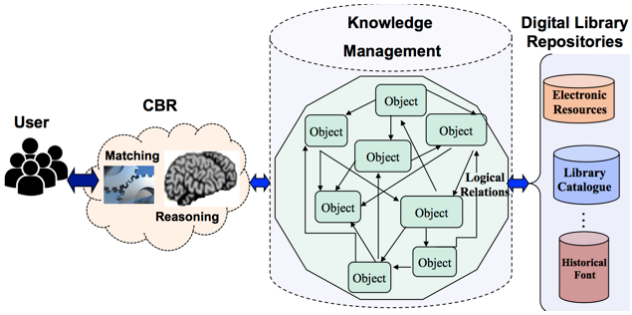


Figure 2. System architecture of OntoSDL

B. The Search Engine

Inference engine contains a CBR component that automatically searches for similar queries-answer pairs based on the knowledge that the system extracted from the questions text [24]. Case Base has a memory organization interface that assumes that whole case-base can be read into memory for the CBR to work with it. Also, we have implemented a new interface, which allows retrieving cases enough to satisfy a SQL query. We used a CBR shell, software that can be used to develop several applications that require cased-based reasoning methodology. We analysed the CBR object-oriented framework development environments JColibri [25]. This framework work as open software development environment and facilitate the reuse of their design as well as implementations. The CBR engine uses an evaluation function to calculate the new case ranking, and the answered question updates the query and the rankings in the displays. The questions are ranked according to their potential for retrieval and matching.

C. The Intelligent User Interface

The acceptability of a system depends to a great extent on the quality of this user interface component [26]. Advanced conversational user interface interacts with users to solve a query, defined as the set of questions selected and answered by the user during conversation. Interface is designed and developed to improve communication between humans and the platform. Interfaces are provided for browsing, searching and facilitating Web contents and services. Interface enhances the flexibility, usability, and

power of human-computer interaction for all users. In realizing the user interface we have exploited knowledge of users, tasks, tools, and content, as well as devices for supporting interaction within different contexts of use. In our system, the user interacts with the system to fill in the gaps to retrieve the right cases. During each search the user selects one item from two displays: ranked questions and ordered cases.

The interfaces provide for browsing, searching and facilitating Web contents and services. It consists of one user profile, consumer search agent components and bring together a variety of necessary information from different user's resources. The user interface helps to user to build a particular profile that contains his interest search areas in the DL domain. The objective of profile intelligence has focused on creating of user profiles: Staff, Alumni, Administrator, and Visitor.

We have developed a graphical selection interface as illustrated in Fig. 3.

Figure 3. User profiles interface

In an intelligence profile setting, people are surrounded by intelligent interfaces merged. Rather than building static user profiles, contextual systems try to adapt to the user's current search. OntoSDL monitors user's tasks, anticipates search-based information needs, and proactively provide users with relevant information. Thus creating a computing-capable environment with intelligent communication and processing available to the user by means of a simple, natural, and effortless human-system interaction. The user enters query commands and the system asks questions during the inference process. Besides, the user will be able to solve new searches for which he has not been instructed, because the user profiles what he has learnt during the previous searchers.

A technical administrator will have a view very different from an end user providing content as an author. Different conceptions, again, will emerge from the perspectives of a digital content aggregator, a 'meta user' or a policy maker. It consists of one user profile, consumer search agent components and bring together a variety of necessary information from different user's resources. Interoperability concepts differ substantially from those of a content consuming end user.

V. CASE-BASED REASONING INTELLIGENT TECHNIQUE

CBR is widely discussed in the literature as a technology for building information systems to support knowledge management, where metadata descriptions for characterizing knowledge items are used. CBR is a problem solving paradigm that solves a new problem, in our case a new search, by remembering a previous similar situation and by reusing information and knowledge of that situation. A new problem is solved by retrieving one or more previously experienced cases, reusing the more similar case, revising, and retaining the case. In our CBR application, problems are described by metadata concerning desired characteristics of a library resource, and the result to a specific search is a pointer to a resource described by metadata. These characterizations are called cases and are stored in a case base. CBR case data could be considered as a portion of the knowledge (metadata) about an OntoSDL object. Every case contains both a solution pointers and problem description used for similarity assessment. Description of the framework case, which is formally described in terms of framework domain taxonomy they are used for indexing cases. The possible solutions described by means of framework instantiation actions and additional information to justifies these steps. The following processes may describe a CBR cycle (Fig. 4):

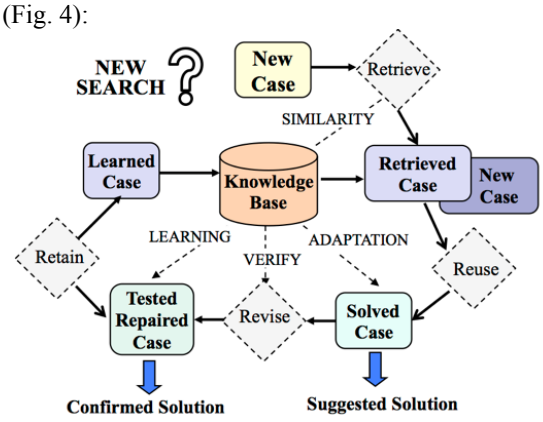


Figure 4. User profiles interface

- Retrieval: main focus of methods in this category is to find similarity between cases. Similarity function can be parameterized through system configuration.
- Reuse: a complete design where case-based and slot-based adaptation can be hooked is provided.
- Revise the proposed solution if necessary. Since the proposed result could be inadequate, this process can correct the first proposed solution.
- Retain the new solution as a part of a new case. This process enables CBR to learn and create a new solution that should be added to the knowledge base.

A. CBR Structure

The development of a quite simple CBR application already involves a number of activities. The actions consist on collecting case and background knowledge, modeling a suitable case representation, defining an accurate similarity measure, implementing retrieval functionality, and implementing user interfaces. Compared with other AI approaches CBR allows to reduce the effort required for knowledge acquisition and representation significantly. This aspect is certainly one of the major reasons for the commercial success of CBR applications. Nevertheless, implementing a CBR application from scratch remains a time-consuming software engineering process and requires a lot of specific experience beyond pure programming skills.

Although CBR claims to reduce the effort required for developing knowledge-based systems substantially compared with more traditional AI approaches. The implementation of a CBR application from scratch is still a time consuming task. We present a novel, freely available tool for rapid prototyping of CBR applications. CBR object-oriented framework development environments JColibri have been used in this study. By providing easy to use model generation, data import, similarity modeling, explanation, and testing functionality together with comfortable graphical user interfaces. The tool enables even CBR novices to rapidly create their first CBR applications. Nevertheless, at the same time it ensures enough flexibility to enable expert users to implement advanced CBR applications [27].

jColibri is and open source framework and their interface layer provides several graphical tools that help users in the configuration of a new CBR system. Our motivation for choosing this framework is based on a comparative analysis between it and other frameworks, designed to facilitate the development of CBR applications. jColibri enhances the other CBR shells: CATCBR, CBR*Tools, IUCBRF, Orenge. Another decision criterion for our choice is the easy ontologies integration. jColibri affords the opportunity to incorporate ontology in the CBR application to use it for case representation and content-based reasoning methods to assess the similarity between them.

B. Retrieval of similar cases process

The main purpose of establishing intelligent retrieval ontology is to provide consistent and explicit metadata in the process of knowledge retrieval. CBR systems typically apply retrieval and matching algorithms to a case base of past search-result pairs. CBR is based on the intuition that new searches are often similar to previously encountered searches, and therefore, that past results may be reused directly or through adaptation in the current situation. Our system provides multilayer retrieval methods:

1. Intelligent profiles interface: Low-level selection of query profile options, which mainly include the four kinds of user. These users can specify certain initial items, i.e., the characteristics and conditions for a search. For this a statistical analysis has been done to determine the importance values and establishing specified user requirements. User searches are monitored by capturing information from different user profiles. This statistical analysis even can in fact lay the foundation for searches in a particular user profile.

2. Ontology semantic search can query on classes, subclasses or attributes of knowledge base, and matched cases are called back.

3. The retrieval process identifies the features of the case with the most similar query. Our inference engine contains the CBR component that automatically searches for similar queries-answer pairs based on the knowledge that the system extracted from the questions text. The system uses similarity metrics to find the best matching case. Similarity measures used in CBR are of critical importance during the retrieval of knowledge items for a new query. Similarity retrieval expands the original query conditions, and generates extended query conditions, which can be directly used in knowledge retrieval. Unlike in early CBR approaches, the recent view is that similarity is usually not just an arbitrary distance measure, but function that approximately measures utility.

We used a computational based retrieval, where numerical similarity functions are used to assess and order the cases regarding the query. The retrieval strategy used in our system is nearest-neighbor technique. This approach involves the assessment of similarity between stored cases and the new input case, based on matching a weighted sum of features. A typical algorithm for calculating nearest neighbor matching is next:

$$similarity(Case_I, Case_R) = \frac{\sum_{i=1}^{n} w_i^{\times sim\left(f_i^I, f_i^R\right)}}{\sum_{i=1}^{n} w_i} \qquad (1)$$

Where w_i is the importance weighting of a feature (or slot), *sim* is the similarity function of features, and f_i^I and f_i^R are the values for feature i in the input and retrieved cases respectively.

The use of structured representations of cases requires approaches for similarity assessment that allow to compares two differently structured objects, in particular, objects belonging to different object classes. An important advantage of similarity-cased retrieval is that if there is no case that exactly matches the user's requirements, this can show the cases that are most similar to his query.

VI. Ontology Design and Development

The main objective of our system is to improve the modelling of a semantic coherence for allowing the interoperability of different modules of environments dedicated to E-Government. We have proposed to use ontology together with CBR in the acquisition of an expert knowledge in the specific domain. The primary information managed in the OntoSDL domain is metadata about institutional resources, such as guides, publications, forms,

digital services, etc. We need a vocabulary of concepts, resources and services for our information system described in the scenario requires definitions about the relationships between objects of discourse and their attributes [28]. OntoSDL project contains a collection of codes, visualization tools, computing resources, and data sets distributed across the grids, for which we have developed a well-defined ontology using RDF language. RDF is used to define the structure of the metadata describing DL resources. Our ontology can be regarded as quaternion OntoSearch:={profile, collection, source, relation), where profiles represent the user kinds. Collection contains all the services and resources of the institutional repository. Source covers the different information suppliers: electronic services, official web pages, publications, guides, etc. Finally, relation element is a set of relationships intended primarily for standardization across ontologies.

We integrated three essential sources to the system: electronic resources, catalogue of documents, and personal Data Base. The W3C defines standards that can be used to design an ontology [29]. We wrote the description of these classes and the properties in RDF semantic markup language. We choose Protégé as our ontology editor, which supports knowledge acquisition and knowledge base development [30]. It is a powerful development and knowledge-modelling tool with an open architecture. Protégé uses OWL and RDF as ontology language to establish semantic relations [31].

Protégé provides an environment for the creation and development of underlying semantic knowledge structures-ontologies and semantically annotated web services. Protégé organizes these elements like a dynamic process workflow. For the construction of the ontology of our system, we followed steps detailed below.

1) Determine the domain and scope of the ontology. This should provide the location of different on-line resources. These are included from different sources: Publications Catalogue, Web Sites, Electronic Resources, etc. Also ontology must be adapted to needs of user kinds.

2) Enumerate important terms in ontology. It is useful to write down a list of all terms we would like either to make statements about or to explain to a user. Initially, it is important to get a comprehensive list of terms without worrying about overlap between concepts they represent, relations among the terms, or any properties that the concepts may have, or whether the concepts are classes or slots.

3) Define the classes and the class hierarchy. When designing the ontology, we first need to group together related resources of the institutional repositories. There are three major groups of resources: users, services, and resources. In order to realize ontology-based intelligent retrieval, we need to build case base of knowledge with inheritance structure. The ontology and its sub-classes are established according to the taxonomies profile. A detailed picture of our effort in designing this ontology is available in Fig. 5. This shows the high level classification of classes to group together OntoSDL resources as well as things that are related with these resources. Profile ontology includes several attributes like Electronic_Resources, Digital_Collections, Publication Catalogue, Public Services, etc.

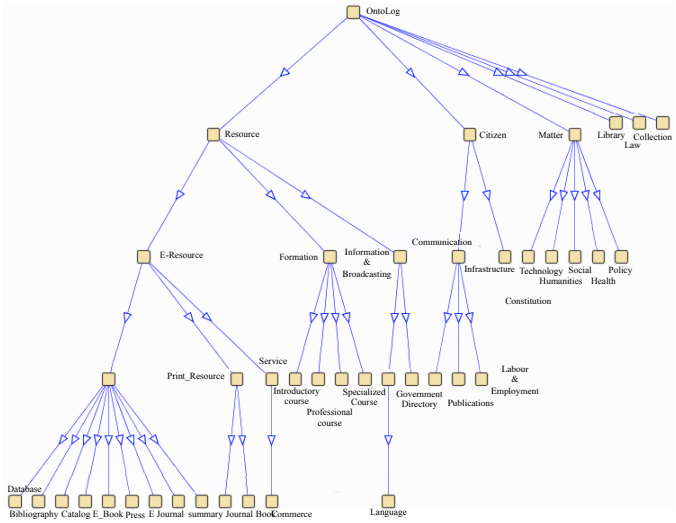


Figure 5. Class hierarchy for the OntoSDL ontology

4) Define the properties of classes and define the facets of the slots. The classes alone will not provide enough information to answer the semantic searches. Once we have defined some of the classes, we must describe the internal structure of concepts. In order to relate ontology classes to each other, we defined our own meaningful properties for the ontology. For this reason, we defined a class hierarchy associated with meaningful properties. Slots can have different facets describing the value type, allowed values, the number of the values (cardinality), and other features of the values the slot can take. In the following, we give a short RDF description that defined the concept of the user teacher that is a subclass of Member_Community_University.

```
<rdf:Description rdf:about="#Teacher">
  <rdfs:comment rdf:datatype=
  "http://www.w3.org/2001/XMLSchema#string">
    Teacher profile for affiliated colleges
  </rdfs:comment>
<rdfs:subClassOf rdf:resource=
    "#Members of the University community"/>
  <rdf:type rdf:resource=
  "http://www.w3.org/2002/07/owl#Class"/>
</rdf:Description>
```

5) Generating the ontology instances with SW languages. To provide a conversational CBR system to retrieve the requested metadata satisfying a user query we need to add enough initial instances and item instances to

knowledge base. The last step is creating individual instances of classes in the hierarchy. Defining an individual instance of a class requires choosing a class, creating an individual instance of that class, and filling in the slot values.

After designing the ontology, we wrote the description of these classes and the properties in RDF semantic markup language. For this purpose, we have followed next steps. First, we choose a certain item, and create a blank instance for item. Then the domain expert, in this case administrative staff fills blank units of instance according the domain knowledge [32]. 11.000 cases were collected for user profiles and their different resources and services. This is sufficient for our proof-of-concept demonstration, but would not be sufficiently efficient to access large resource sets. Each case contains a set of attributes concerning both metadata and knowledge.

However, our prototype is currently being extended to enable efficient retrieval directly from a database, which will enable its use for large-scale sets of resources. As a plus, domain specific rules defined by domain experts (manually or by tools) can infer more complex high-level semantic descriptions, for example, by combining low-level features in local repositories. On one hand, the rules can be used to facilitate the task of resource annotation by deriving additional metadata from existing ones.

Keeping in mind that our final goal is to reformulate queries in the ontology to queries in another with least loss of semantics, we come to a process for addressing complex relations between two ontologies. As mentioned in previous sections, relations among ontologies can be composed as a form of declarative rules, which can be further handled in inference engines. In our approach, we choose to use the Semantic Web Rule Language (SWRL), which is based on a combination of OWL DL and OWL Lite with the case-based reasoning sublanguages, to compose declarative search rules [33].

VII. EXPERIMENTAL EVALUATION

Experiments have been carried out in order to test the efficiency of AI and ontologies in retrieval information in a DL. These are conducted to evaluate the effectiveness of run-time ontology mapping. The main goal has been to check if the mechanism of query formulation, assisted by an agent, gives a suitable tool for augmenting the number of significant documents, extracted from the DL to be stored in the CBR. The user begins the search devising the starting query. Suppose the user is looking for some resource about "Computer Science electronic resource" in the library digital domain of Seville (Fig. 6).

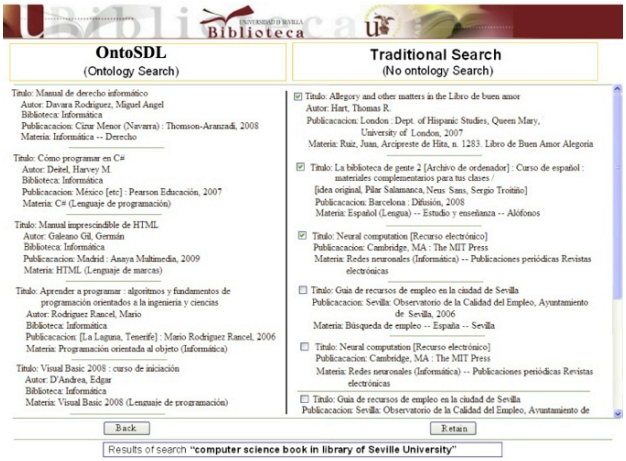


Figure 6. Search engine results page

The user inputs the keywords in the user profile interface. The required resources should contain some knowledge about "Computer Science" and related issues. After searching, some resources are returned as results. The results include a list of web pages with titles, a link to the page, and a short description showing where the keywords have matched content within the page.

We have compared our prototype with some semantic search engines like Hakia, Lexxe, SenseBot, etc. However, we have focused in Google because is the world's dominant search engine and Google has made significant inroads in semantic indexing in search. It is a fact that deep inside Google is based on breakthrough semantic search techniques that are transforming Google's search results [34].

For our experiments, we considered 50 users with different profiles. Therefore, we could establish a context for the users, they were asked to at least start their essay before issuing any queries to OntoSDL. They were also asked to look through all the results returned by OntoSDL before clicking on any result. We compared the top 10 search results of each keyword phrase per search engine. Our application recorded which results on which they clicked, which we used as a form of implicit user relevance in our analysis. We must consider that retrieved documents relevance is subjective. That is different people can assign distinct values of relevance to a same document.

In each experiment, we report the average rank of the user-clicked result for our baseline system, Google and for our search engine OntoSDL. In our study, we have agreed different values to measure the quality of retrieved documents, excellent, good, acceptable and poor. Next, we calculated the rank for each retrieval document by combining the various values and comparing the total number of extracted documents and documents consulted by the user (Table 1).

TABLE I. ANALYSIS OF RETRIEVED DOCUMENTS RELEVANCE FOR SELECT QUERIES

	Excellent	Good	Acceptable	Poor
OntoSDL	7,50%	41,50%	40,60%	10,40%
Google	2,60%	27,90%	43,40%	26,10%

After the data was collected, we had a log of queries averaging 5 queries per user. Of these queries, some of them had to be removed, either because there were multiple results clicked, no results clicked, or there was no information available for that particular query. The remaining queries were analyzed and evaluated. These results are presented in Fig. 7.

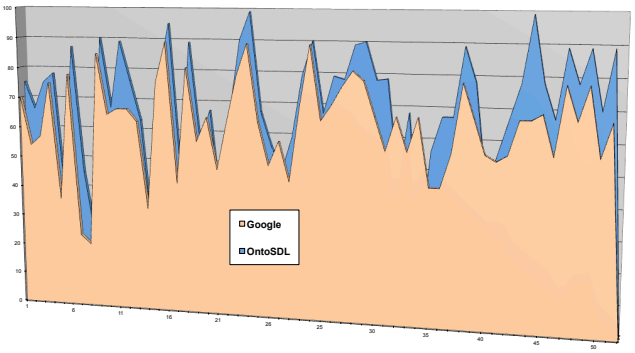


Figure 7. Search engine results page

In the digital library domain we can observe that best final ranking was obtained for our prototype. OntoSDL achieves an interesting improvement over the performance of Google. Other significance test is the analysis of the number of searches that have been resolved satisfactory by OntoSDL. As noted in Table I, our system performs satisfactorily with about a 91.6% rate of success in real cases.

Another important aspect of the design and implementation of an intelligent system is determination of the degree of speed in the answer that the system provides. During the experimentation, heuristics and measures that are commonly adopted in information retrieval have been used. While the users were performing these searches, an application was continually running in the background on the server, and capturing the content of queries typed and the results of the searches. Statistical analysis has been done to determine the importance values in the results. We can establish that speed in our system improves the proceeding time and the average of the traditional search engine. The results for OntoSDL are 9.15% better than proceeding time and 11.9% better than executing time searches/sec in the traditional search engines.

VIII. CONCLUSION AND FUTURE WORK

We have investigated how semantic technologies and AI can be used to provide additional semantics from existing resources in digital libraries. We described an effort to design and develop a prototype for management the resources in a library such as OntoSDL project, and to exploit them to aid users as they select resources. Our study addresses the main aspects of a Semantic Web knowledge retrieval system architecture trying to answer the requirements of the next-generation Semantic Web user. This scheme is based on the next principle: knowledge items are abstracted to a characterization by metadata description and it is used for further processing.

For this purpose, we presented a system based in ontology and AI architecture for knowledge management in the Seville DL. First of all, to put our aims into practice, we should develop the domain ontology and study how the content-based similarity between the concepts typed attributes could be assessed in CBR system. A dedicated inference mechanism is used to answer queries conforming to the logic formalism and terms defined in our ontology. We have been working on the design of entirely ontology-based structure of the case and the development of our own reasoning methods in jColibri to operate with it. It introduced a prototype web-based CBR retrieval system, which operates on an RDF file store. Furthermore, an intelligent agent was illustrated for assisting the user by suggesting improved ways to query the system on the ground of the resources in a DL according to his own preferences, which come to represent his interests.

Finally, the study analyses the implementation results, and evaluates the viability of our approaches in enabling search in intelligent-based digital repositories. OntoSDL can be part of a bigger framework of interacting global information networks including e.g., other digital libraries, scientific repositories and commercial providers. The framework relies as much as possible on standards and existing building blocks as well as is based on web standards.

The results demonstrate that by improving representation by incorporating more metadata from within the information and the ontology into the retrieval process, the effectiveness of the information retrieval is enhanced. Future work will concern the exploitation of information coming from others institutional repositories and digital services. Furthermore, we propose refine the suggested queries, to extend the system to provide another type of support, as well as to refine and evaluate the system through user testing. It is also necessary the development of an authoring tool for user authentication, efficient ontology parsing and real-life applications.

REFERENCES

[1] A. Martín and C. León, “Intelligent Technique to Accomplish a Effective Knowledge Retrieval from Distributed Repositories,” in Proc. Third International Conference on Intelligent Systems and Applications (INTELLI), pp. 97-102, Seville, Spain, 2014.

[2] A. Jimeno-Yepes, R. Berlanga-Llavori, and D. Rebholz-Schuhmann, “Ontology refinement for improved information retrieval,” Information Processing & Management, Volume 46 (Issue 4), Semantic Annotations in Information Retrieval, 2010.

[3] M.C. Diaz-Galiano, M.T. Martin-Valdivia, and L.A. Urena-Lopez, "Query expansion with a medical ontology to improve a multimodal information retrieval system," Computers in Biology and Medicine, Volume 39 (Issue 4), pp. 96-403, 2009.

[4] L. Chen, "Design and implementation of intelligent library system," Library Collections, Acquisitions, and Technical Services, Volume 32 (Issues 3–4), pp.127-141, 2008

[5] J.S. Cho and K.H. Hyun, "Meta-ontology for automated information integration of parts libraries," Computer-Aided Design, Volume 38 (Issue 7), pp. 713-725, 2006.

[6] H. Sasaki and Y.A. Kiyoki, "A formulation for patenting content-based retrieval processes in digital libraries," Information Processing & Management, Volume 41 (Issue 1), pp. 57-74, 2005.

[7] D. Bainbridge, M. Dewsnip, and I.H. Witten, "Searching digital music libraries." Information Processing & Management, Volume 4, (Issue 1), pp. 41-56, 2005.

[8] C.M. Toledo, M.A. Ale, O. Chiotti, and M.R. Galli, "An Ontology-driven Document Retrieval Strategy for Organizational Knowledge Management Systems," Electronic Notes in Theoretical Computer Science (Vol. 281), pp. 21-34, 2011.

[9] D. Govedarova, S. Stoyanov, and I. Popchev, "An Ontology Based CBR Architecture for Knowledge Management in BULCHINO Catalogue," in Proc. International Conference on Computer Systems and Technologies (CompSysTech), 2008.

[10] P. Warren, "Applying semantic technologies to a digital library: a case study", Library Management Journal, Emerald, 2005.

[11] H. Stuckenschmidt and F.V. Harmelen, "Ontology-based metadata generation from semi-structured information," K-CAP, pp. 163–170, ACM, 2011.

[12] J. Toussaint and K. Cheng, "Web-based CBR (case-based reasoning) as a tool with the application to tooling selection," International Journal of Advanced Manufacturing Technology, 2006.

[13] GAIA - Group for Artificial Intelligence Applications. *jCOLIBRI project - Distribution of the development environment*, [Online]. Available from: http://gaia.fdi.ucm.es/research/colibri/jcolibri/ 2015.04.25

[14] Y. Sure and R. Studer, "Semantic Web technologies for digital libraries," Library Management Journal, Emerald, Vol. 26, pp. 190-195, 2005.

[15] I.H. Witten and D. Bainbridge, "How to Build a Digital Library" Morgan Kaufmann, 2003.

[16] H. Ding, "Towards the metadata integration issues in peer-to-peer based digital libraries," GCC. H. Jin, Y. Pan, N. Xiao, and J. Sun, (eds.) (LNCS), Vol. 3251, Berlin, Germany, Springer, 2004.

[17] R. Guha, R. McCool, and E. Miller, "Semantic search," In Proceedings of WWW2003, 2003.

[18] G.F. Luger, "Artificial Intelligence, Structures and Strategies for Complex Problem Solving," 4ª edition. Ed. Pearson Education Limited, 2002.

[19] Z. Sun and G. Finnie, "Intelligent Techniques in E-Commerce: A Case-based Reasoning Perspective," Heidelberg: Springer-Verlag, 2004.

[20] SEC. Commission Staff Workking Paper: linking up Europe, *the importance of interoperability for egovernment services*, [Online]. Available from: http://europa.eu.int/ISPO/ida/export/files/en/1523.pdf, 2015.05.3

[21] MAP. Aplicaciones utilizadas para el ejercicio de potestades. Criterios de Seguridad, Normalización y Conservación. *Ministerio de Administraciones Públicas*. [Online]. Available from: http://www.csi.map.es/csi/criterios/index.html, 2014.03.05

[22] EIF. *European Interoperability Framework Version 2*. [Online]. Available from: http://ec.europa.eu/isa/strategy/doc/annex_ii_eif_en.pdf, 2015.04.19.

[23] S. Staab and R. Studer, "Handbook on Ontologies," International Handbooks on Information Systems, Springer, Berlin, 2005.

[24] M. Bridge, H. G¨oker, L. McGinty, and B. Smyth, "Case-based recommender systems," Knowledge Engineering Review, 2006.

[25] B. Díaz-Agudo, P.A. González-Calero, J.Recio-García, and A Sánchez-Ruiz, "Building CBR systems with jColibri," Journal of Science of Computer Programming, Volume 69, Issues 1–3, 1 December 2007, pp. 68–75, doi: dx.doi.org/10.1016/j.scico.2007.02.004.

[26] D. Quan and D.R. Karger, "How to make a semantic web browser," I Proc. of Thirteenth International World Wide Web Confecence (WWW), pp. 17–22, New York, New York, USA, Vol. 12, Issue 1, pp. 1- 40, 2004.

[27] L.A. Breslowm and D.W. Aha, "Simplifying decision trees: A survey," The Knowledge Engineering Review archive, Cambridge University Press New York, NY, USA, 1997.

[28] D. Taniar And J.W. Rahayu, "Web semantics and ontology," Hershey, PA: Idea Group, 2006.

[29] W3C. *RDF Vocabulary Description Language 1.0: RDF Schema*. [Online]. Available from: http://www.w3.org/TR/rdf-schema/, 2015.02.10.

[30] PROTÉGÉ. *The Protégé Ontology Editor and Knowledge Acquisition System*. [Online]. Available from: http://protege.stanford.edu/, 2015.04.05.

[31] J. Heflin, "OWL Web Ontology Language Use Cases and Requirements," W3C Recommendation, 2004.

[32] M. Horridge and H. Knublauch, "A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools," The University Of Manchester, United Kingdom, 2004.

[33] S. Bechhofer, F.V Harmelen, J. Hendler, I. Horrocks, D.L. McGuinness, P.F Patel-Schneider, and L.A. Stein, "OWL web ontology language reference", W3C recommendation. Volume 10 (Issue February), Publisher W3C, 2004.

[34] D. Amerland, "Google Semantic Search: Search Engine Optimization (SEO) Techniques That Get Your Company More Traffic, Increase Brand Impact and Amplify Your Online Presence", Que Publishing Kindle Edition, July, 2013.

Fuzzy Control for Gaze-Guided Personal Assistance Robots: Simulation and Experimental Application

Carl A. Nelson
Department of Mechanical and Materials Engineering
University of Nebraska-Lincoln
Lincoln, NE, USA
email: cnelson5@unl.edu

Xiaoli Zhang, Jeremy Webb, and Songpo Li
Department of Mechanical Engineering
Colorado School of Mines
Golden, CO, USA
email: xlzhang@mines.edu, jewebb@mymail.mines.edu, soli@mines.edu

***Abstract*—As longer lifespans become the norm and modern healthcare allows individuals to live more functional lives despite physical disabilities, there is an increasing need for personal assistance robots. One of the barriers to this shift in healthcare technology is the ability of the human operator to communicate his/her intent to the robot. In this paper, a method of interpreting eye gaze data using fuzzy logic for robot control is presented. Simulation results indicate that the fuzzy logic controller can successfully infer operator intent, modulate speed and direction accordingly, and avoid obstacles in a target following task relevant to personal assistance robots. The fuzzy logic approach is then validated through navigation experiments using a small humanoid robot.**

Keywords—gaze tracking; fuzzy logic; autonomous robot; obstacle avoidance; personal assistance robot

I. INTRODUCTION

With lifespans increasing worldwide due to advancements in healthcare and related technologies, the importance of care for the elderly and disabled is increasing. In particular, there is a shifting emphasis in technology development towards improving quality of life in the face of diminishing physical capabilities. One of the burgeoning areas of this trend is personal assistance robotics. In a typical scenario, a robot assistant may be present in the home to help with basic day-to-day tasks (e.g., object retrieval), especially those tasks requiring navigation throughout the home, since age- or disability-related mobility limitations may keep an individual from performing all these tasks personally. In extreme circumstances, it can even be challenging to give instructions to the robotic assistant, as in the case where the individual is not physically able to type, speak, or otherwise provide clear inputs to the human-robot interface. Here, we build on work presented in [1] and present progress towards a robotic assistance system which relies on gaze tracking, including eye blinking patterns, to infer a person's intent and thereby create instructions for the robot. In this paper, we specifically focus on the intelligent inference of intent based on gaze and blinking input.

This problem is an extension of the task of robotic target following and path planning. Significant work has been done in this area of service robotics, where a robot is to follow a moving target. For instance, some have used computer vision, using optical flow algorithms to track the target [2][3]. Other computer vision-based approaches have used Kalman filters for improving the accuracy of tracking [4]. Other tracking methods include the use of depth images with verification via a state vector machine [5], or following acoustic stimuli [6]. Control approaches in these target-following scenarios include potential field mapping [7] and a variety of other techniques. Of particular interest are fuzzy logic controllers [6][8][9], which tend to be used primarily for steering, but can easily be adapted to handle various types of linear and nonlinear systems [10]. Here, we will describe a fuzzy logic controller which not only determines the robot's heading based on the location of the target, but also avoids obstacles and modulates speed based on the perception of intent from the combined gaze direction and blink frequency inputs. The authors believe this perception of intent combined with heading, speed, and obstacle avoidance to be unique with respect to the state of the art in robot guidance. This is conceptually based in part on recent work demonstrating how such a combined input using operator gaze could be used for automatic control of endoscope positioning in surgical tasks [11] using a commercially available eye tracking system, which is also similar to the work described in [12]. Existing examples of robot control using gaze input are relatively scarce. In [13][14][15], specially identified eye movements, such as looking up, down, left and right, were mapped to wheelchair steering commands to drive it forward, backward, left and right. In [16], on-screen buttons were created to activate joint rotation of an articulated robot arm such that a user could steer the arm by gazing at the buttons. However, steering a robot arm in this manner is very inefficient and can be exhausting for the user, as he/she has to explicitly control every movement of the arm. In [17], the user's gaze vector was estimated and served as a pointing line along which the robot could search to retrieve the first found object. However, as the exact location of the object was not calculated, extensive searching had to be carried out along the gaze vector to locate the object. The gaze-based robot control approach proposed in this article extends beyond the most typical uses of eye gaze, which tend to be for two-dimensional human-computer interfaces [18], to interacting in the three-dimensional context using gaze tracking for activities of daily living, i.e., using personal assistance robots. Although the robotic assistance scenario clearly would involve more subtasks (such as object manipulation), we limit our treatment in this paper to development of controllers which modulate robot heading and speed while avoiding obstacles, for navigating in a potentially cluttered environment using gaze as the input data stream.

The remainder of this paper is organized as follows. In Section II, the eye gaze data and the fuzzy logic controller are described. In Section III, simulation results are presented, followed by experimental validation using a small humanoid robot platform. Section IV includes conclusions and recommendations for future work.

II. METHODS

A. Test Dataset and Simulation

A gaze dataset was artificially generated to have spatiotemporal characteristics similar to those described in [11], in a planar workspace. The data were arbitrarily assumed to be sampled at 10 Hz and included a logical *blink* data channel in addition to the *x* and *y* gaze target channels on the interval [-0.5 0.5], providing a total of over 23 seconds of simulated robot tracking. Due to the noisy nature of gaze data, the target *X* was determined by a linear weighted average of the previous *n* data points *P*, with *n* = 20:

$$X_k = \frac{2}{n}\sum_{i=k-n}^{k}\left(1-\frac{k-i}{n}\right)P_i. \quad (1)$$

In this particular dataset, there are five intended target locations, characterized by dwelling gaze and higher blink frequency, and it is assumed that a supplementary action such as object placement or retrieval would follow target acquisition (although this supplementary action is beyond the scope of this preliminary study). Within the workspace, three round obstacles were defined to test the ability of the simulated robot to avoid obstacles while seeking a target. The data were imported into MATLAB (The MathWorks, Natick, MA) for simulation of gaze-based robotic target tracking.

B. Fuzzy Logic Controller

A Mamdani-type fuzzy logic controller [19] with five inputs and three outputs was created using the Fuzzy Logic Toolbox in MATLAB; the Mamdani-type model handles multi-input, multi-output (MIMO) problems better than the Sugeno-type alternative. The inputs, shown in Table I, were intended to take into account the control objectives: to track a target at an appropriate speed based on uncertain data while avoiding obstacles. Distance to the target is captured by target Δx and target Δy, the degree of uncertainty of the target's position is expressed by the target variability, and the presence of obstacles in the path from the robot's position to the target is quantified by the obstacle distance. The blink frequency is used to capture operator intent and desired speed. The outputs, also shown in Table I, were used to control the speed and heading of the robot, including steering adjustments for obstacle avoidance. All of the membership functions were triangular, as shown in Fig. 1, and their parameters were tuned by hand using a minimal amount of trial and error.

The target was determined using a weighted average of the gaze data as in (1), the target and obstacle distance variables were then calculated using the Pythagorean theorem, and target variability was represented by the standard deviation of the gaze input data over the averaging window. (It is noteworthy that target variability is likely to be the input parameter most sensitive to individual characteristics, and therefore would need to be tuned for each individual's gaze "signature." In this case, it was tuned to accommodate the characteristics of the dataset described in Section II.A.)

TABLE I. FUZZY CONTROLLER VARIABLES AND THEIR TRIANGLUAR MEMBERSHIP FUNCTIONS EXPRESSED IN MODAL FORM [LOWER BOUND, MODE, UPPER BOUND]

Input/ Output	Variables		
	Name	***Units***	***Membership Functions***
I	Target Δx	distance	negative [-1, -0.5, 0] zero [-0.1, 0, 0.1] positive [0, 0.5, 1]
I	Target Δy	distance	negative [-1, -0.5, 0] zero [-0.1, 0, 0.1] positive [0, 0.5, 1]
I	Target variability	distance	zero [-0.1, 0, 0.1] low [0.05, 0.25, 0.45] high [0.35, 1, 1.4]
I	Blink frequency (normalized)	-	zero [-0.4, 0, 0.4] low [0.1, 0.5, 0.9] high [0.6, 1, 1.4]
I	Obstacle distance	distance	zero [-0.2, 0, 0.2] low [0, 0.3, 0.6] high [0.35, 1, 1.4]
O	Speed	distance/ time	zero [-0.4, 0, 0.4] low [0.1, 0.5, 0.9] high [0.6, 1, 1.4]
O	Heading	rad	up [0.125, 0.25, 0.375] up/right [0, 0.125, 0.25] right [-0.125, 0, 0.125] down/right [0.75, 0.875, 1] down [0.625, 0.75, 0.875] down/left [0.5, 0.625, 0.75] left [0.375, 0.5, 0.625] up/left [0.25, 0.375, 0.5]
O	Heading adjustment	rad	zero [-0.4, 0, 0.4] low [0.1, 0.5, 0.9] high [0.6, 1, 1.4]

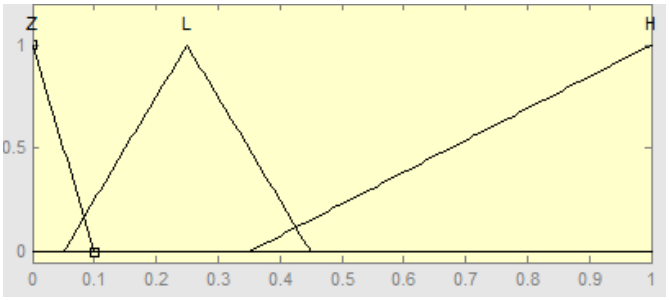


Figure 1. Membership functions for target variability (zero, low, and high).

Blink frequency was normalized to the interval [0 1] by assuming that four blink events within the 20-sample averaging window was high (achieving a value of 1), and lower blinking rates in the same window of time receive a proportionally smaller membership value. If no obstacles were detected in the direct path between the robot and target, the obstacle distance was set to its maximum value of 1. The other distance-based variables did not need to be explicitly normalized since the workspace was set up as a unit square. It should also be noted (referring to Table I) that in certain cases (e.g., the "zero" membership functions for target variability, obstacle distance, and blink frequency), negative values (which do not have physical meaning) were used in order to create membership

functions for which the lower bound is also the mode, without causing errors in the software.

Concerning the output variables, the maximum speed was constrained to a value of 0.25 (covering one-fourth the workspace in one second at maximum speed), and the maximum heading adjustment for obstacle avoidance was set at ±100°. The heading variable was scaled to allow the robot to steer within the full 360° range.

Fifteen rules were defined to characterize the influence of the five input variables on the three outputs. In particular, four rules capture the influence of the inputs on the output variable *speed*, eight rules accommodate the division of heading into eight regions in polar coordinates, and the remaining three rules govern obstacle avoidance. The rules defining the fuzzy logic controller are as follows:

1. IF *blink* IS *high* THEN *speed* IS *high*
2. IF *target* Δ*x* IS *positive* OR *target* Δ*x* IS *negative* OR *target* Δ*y* IS *positive* OR *target* Δ*y* IS *negative* THEN *speed* IS *high*
3. IF *target* Δ*x* IS *zero* AND *target* Δ*y* IS *zero* THEN *speed* IS *zero*
4. IF *target variability* IS *high* OR *blink* IS *low* THEN *speed* IS *low*
5. IF *target* Δ*x* IS *positive* AND *target* Δ*y* IS *zero* THEN *heading* IS *right*
6. IF *target* Δ*x* IS *positive* AND *target* Δ*y* IS *positive* THEN *heading* IS *up/right*
7. IF *target* Δ*x* IS *positive* AND *target* Δ*y* IS *negative* THEN *heading* IS *down/right*
8. IF *target* Δ*x* IS *negative* AND *target* Δ*y* IS *zero* THEN *heading* IS *left*
9. IF *target* Δ*x* IS *negative* AND *target* Δ*y* IS *positive* THEN *heading* IS *up/left*
10. IF *target* Δ*x* IS *negative* AND *target* Δ*y* IS *negative* THEN *heading* IS *down/left*
11. IF *target* Δ*x* IS *zero* AND *target* Δ*y* IS *positive* THEN *heading* IS *up*
12. IF *target* Δ*x* IS *zero* AND *target* Δ*y* IS *negative* THEN *heading* IS *down*
13. IF *obstacle distance* IS *zero* THEN *heading adjustment* IS *high*
14. IF *obstacle distance* IS *low* THEN *heading adjustment* IS *low*
15. IF *obstacle distance* IS *high* THEN *heading adjustment* IS *zero*

The first four rules govern the robot's speed. Higher blink rates imply a more focused operator intent and cause increased speed (rule 1). Conversely, high gaze variability or low blink rate imply a less sure target and lead to lower speed (rule 4). The higher the distance to the target, the higher the necessary speed to reach it in a timely manner, and speed should drop to zero as the target is reached (rules 2-3). It should be noted that lower speeds are sometimes desirable to conserve energy either when the goal is unclear or has been reached.

Rules 5-12 pertain to heading. These are relatively straightforward and use the four cardinal directions and the four semi-cardinal directions to navigate in the planar map based on the relative target distance in the *x* and *y* directions. This can be thought of as a fuzzy calculation of inverse tangent for the heading angle using *target Δx* and *target Δy* as inputs.

The remaining three rules (rules 13-15) constitute the robot's obstacle avoidance behavior. The closer the obstacle, the larger the heading adjustment applied to go around it. Whether this adjustment is added or subtracted from the heading variable is determined by whether the obstacle centroid is to the right or the left of the straight line along the robot's heading.

C. Experiments

In addition to simulation, the fuzzy logic controller was implemented on a commercially available small humanoid robot (NAO, Aldebaran Robotics). This platform was chosen, in part, because it has computer vision and sonar sensors suitable for obstacle detection as a built-in feature, so it is expected to scale well towards more advanced demonstrations in the future. The overall architecture, illuminating the concept and the correlations of components in the system, is shown in Fig. 3. The system contains four parts: an eye tracking system which can track where the user is looking on a monitor, a camera which provides video feedback to the user and functions as a global tracking system, the host system which is responsible for collecting and interpreting the gaze data into robot motion commands and sending these motion commands wirelessly to the robot for navigation, and the NAO robot. A rectangular workspace area (3m x 4m) was defined, with two round obstacles marked, similar to the setup of the simulation experiments.

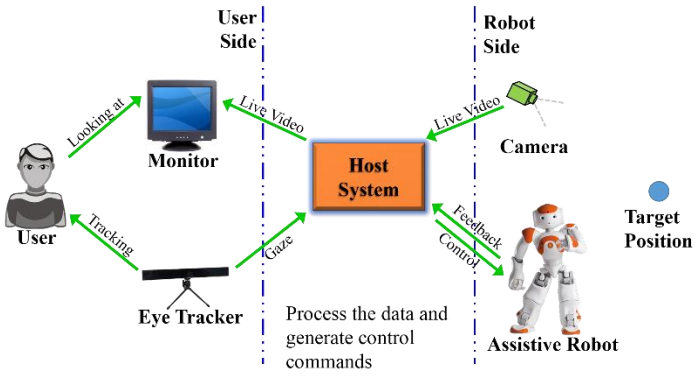


Figure 2. Interaction in the gaze-based system for robot navigation.

In the presented system, the user, sitting in a chair, watches the live video fed from the camera (shown in Fig. 3). The user's gaze on the video is sensed, from which the user's visual attention is detected. The detected visual attention represents

the target position that the user wants the robot to approach. Then the visual attention position as well as the online collected gaze and blink data are interpreted by the fuzzy controller into a series of control commands to generate the optimal trajectory for the robot to approach the target position. The response of the NAO robot was recorded.

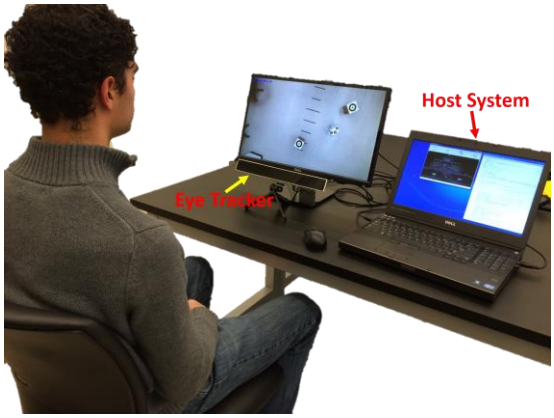


Figure 3. User control interface of the system.

A GP3 Eye Tracker (GazePoint) is used to track where the user is looking on a monitor. GP3 is a video-based remote eye tracking system which allows head movement of a user in a volume of 25×11×30cm^3, without significant degradation of the tracking accuracy. It can report the gaze data at 60 Hz with an accuracy of 0.5˚ - 1˚ and drift of less than 0.3˚. Calibration is required before it can provide accurate eye tracking. (Although the long-term objective is to track 3D gaze rather than planar gaze on a screen, and hardware is under development to achieve such 3D gaze tracking, the work presented here is focused on validation of the control concept, and as such, planar gaze tracking consistent with the simulation approach presented in Section II.A-B is appropriate for these experiments.)

Human eyes are capable of making many different movements. Some are involuntary, such as rolling, nystagmus, drift or microsaccades; these involuntary movements are superimposed on the voluntary eye movements such as gaze fixation. The gaze data estimated from human eye movements will thus include noise from the superimposed involuntary eye movements, and must therefore be filtered to extract the gaze data that is related to the attentional processes of the viewer. An adaptive sliding window filter is employed to eliminate this noise, and a dwell time method is utilized to extract visual attention. Due to the complexity of human eye movement, this filter is slightly more complex than the one used in simulation (Eq. (1)).

The adaptive sliding window filter is illustrated in Eqs. (2) and (3). N is the size of the sliding window. P_i and $\tilde{P}_i$ are the i^{th} gaze point before and after filtering, respectively. E_i is an influence coefficient, calculated using Eq. (3), which indicates the degree of influence a newly received gaze point has on the attention extraction. The influence coefficient of a gaze point is determined by the relative distance from that point to the mean of all the gaze points in the current sliding window. The output of the filter is the mean of all the weighted gaze points. This filter is intended to remove the effects of blinking, attention shifting, and tracking failure. At the same time it can smooth the gaze points by eliminating effects of involuntary eye movements such as rolling, nystagmus, drift and microsaccades.

$$\tilde{P}_i = \frac{1}{\sum_{k=1}^{N} E_{i-k+1}} \left\{ \sum_{j=1}^{N-1} \tilde{P}_{i-j} * E_{i-j} + P_i * E_i \right\} \quad (2)$$

$$E_i = \begin{cases} 1, & \left\| P_i - \frac{1}{\sum_{k=1}^{N} E_{i-k}} \sum_{j=1}^{N} \tilde{P}_{i-j} * E_{i-j} \right\| \leq threshold \\ 0, & \left\| P_i - \frac{1}{\sum_{k=1}^{N} E_{i-k}} \sum_{j=1}^{N} \tilde{P}_{i-j} * E_{i-j} \right\| > threshold \end{cases} \quad (3)$$

Using gaze as an input signal for human-robot interaction requires the differentiation of normal behavioral eye movements and intentional eye "commands," which is known as the Midas touch problem [20]. The "select" or "click" command is usually derived from either blink or gaze dwell time, which is used as a confirmation of a specific command from the eyes. The dwell time method is derived from the fact that a person's eyes stay focused on a target when he/she concentrates on a visual target. In this paper, a dwell time of 2 seconds was used. Once the user stares at an object for more than 2 seconds, the system considers that object as the visual attention point of the user, and this triggers a series of motion commands of the robot.

Due to the challenges in implementing the proposed control method on a real robot, a few changes were made to the simulated control flow. In particular, since the NAO robot travels relatively slowly and it would be cumbersome for a user to have to continuously focus on a target location until the robot reached it, the target location was acquired at the beginning of each movement by fixating on a single location for two seconds. Once the target was known, NAO would first turn towards the target and then begin moving. At this point, the fuzzy rules took over, controlling the heading based on the current location of the robot, and the speed based on the distance from the target, the gaze variability, and the blink rate. Note that the user is not required to focus on the target the whole time so if he/she is looking at many different locations while the robot is moving, the gaze variability will be high and rule 4 above will be invoked. On the other hand, if the user is engaged with the task at hand and focused on the robot or the target, the gaze variability will be low. Additional differences are in the speeds used. The output from the fuzzy logic rules was calculated and sent to the robot at 8 Hz while the gaze data were collected at 60 Hz with an averaging window size of 70 data points.

III. RESULTS

The simulations described in Sections II.A-B and the experiments described in Section II.C generally produced similar outcomes validating the approach. These outcomes are described as follows.

A. Simulation Outcomes

Simulation in MATLAB revealed the ability of the fuzzy logic controller to simultaneously determine human intent from the combined gaze location and blink data, use this intent to modulate robot speed, follow a moving target, and avoid obstacles. In Fig. 4, it can be observed that the robot (whose position is indicated by red diamond markers) can start at a location somewhat removed from the initial target, quickly acquire the target, and then follow it consistently without colliding with obstacles in the workspace. It can be noted that filtering the raw gaze data (black dot markers) smoothes but does not significantly alter the target path (blue circle markers), and that the robot follows the target reasonably closely when it is not engaged in obstacle avoidance.

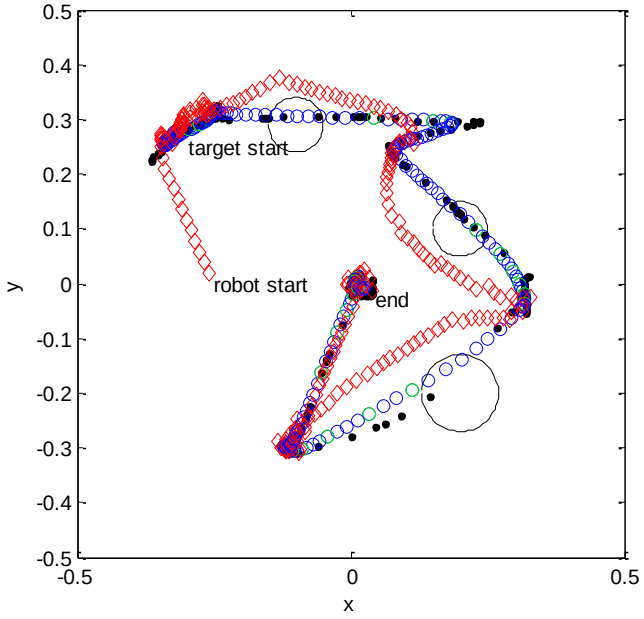


Figure 4. Target following behavior: robot (red diamond markers) follows target (blue circles) while avoiding fixed environmental obstacles. Green circles indicate target location with a *blink* event. Targets of definite interest (based on dwell duration and blink frequency) are at approximately (-0.3, 0.3), (0.1, 0.3), (0.3, 0), (-0.1, -0.3), and (0, 0). Raw gaze data are shown as black dots.

The more interesting outcomes of the simulation are highlighted in Figs. 5-8, in which the input/output model parameters from the simulation of Fig. 4 are plotted separately to elucidate the effects of the fuzzy rule set. In Fig. 5, one can see that robot speed tends to increase with blink frequency, as intended (rules 1 and 4). High speed at low blink value can be attributed to the effects of target distance (particularly at the beginning of the simulation, rule 2). Note that the results in Fig. 5 are striated at discrete levels, since blinking is a discrete, logical event; this could be smoothed by applying an averaging method similar to that used in target determination.

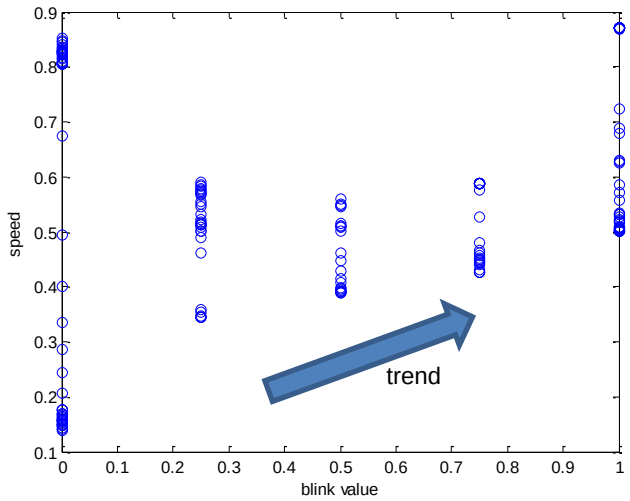


Figure 5. Output speed as a function of *blink* membership function value: a positive correlation is noted.

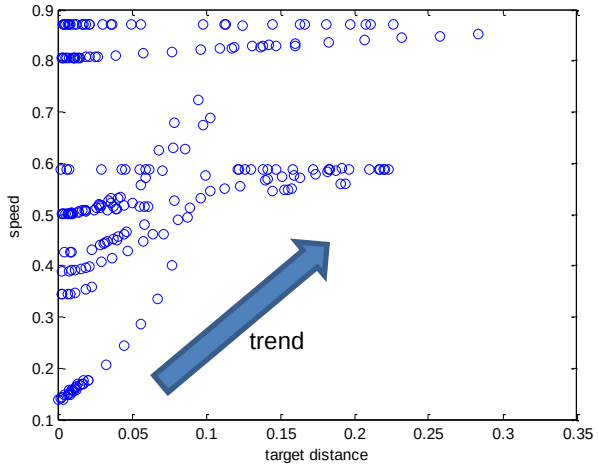


Figure 6. Output speed as a function of target distance: a positive correlation is noted.

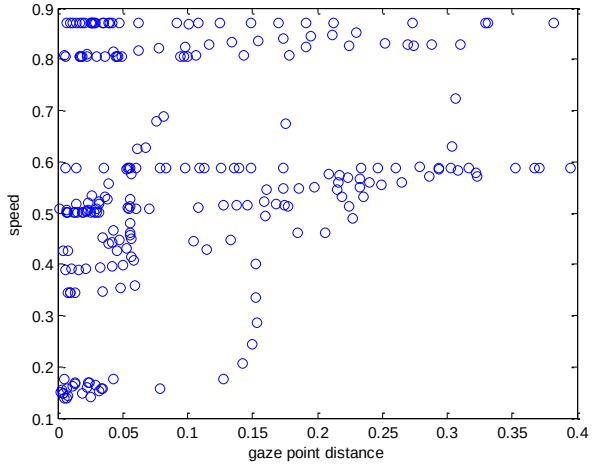


Figure 7. Output speed as a function of distance to current gaze location: correlation is much less pronounced.

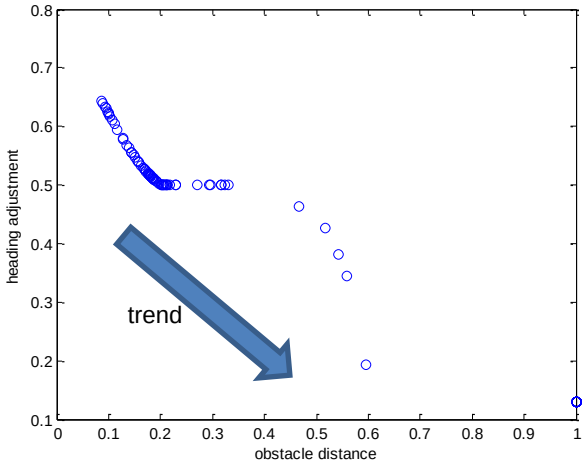


Figure 8. Modulation of heading adjustment based on obstacle distance demonstrates effective obstacle avoidance.

Target distance also has an important effect on speed (rule 2), as shown in Fig. 6. In contrast, Fig. 7 illustrates that the relationship between robot speed and distance from the robot to the actual gaze point is less pronounced, since the target is based on a weighted average of the gaze point and is thus a less noisy signal. The interdependence of speed on multiple input parameters is evident in Figs. 5 and 6. The effectiveness of the obstacle avoidance behavior (rules 13-15) is shown in Fig. 8 by the clean heading adjustment curve. These results illustrate the suitability of the fuzzy controller for satisfying the control objectives noted in Section I.

B. Experimental Outcomes

The experimental setup is shown in Fig. 9. It includes the NAO robot with a unique marker, and two obstacles with a different unique marker, in a rectangular area. The target gaze position is indicated by a small blue square on the figure. Feedback to the control computer (not shown in the figure) is done through overhead camera capture of the scene. In this experiment shown, there are two target locations at which the gaze lingers (indicated in the latter two parts of Fig. 9).

The data recorded in the experiment of Fig. 9 are shown in Fig. 10 as a time lapse, similar to the simulation results of Fig. 4. It is clear that the fuzzy logic controller allows the robot to effectively seek targets while avoiding obstacles, just as in the simulation. Additional evidence of this is illustrated in Figs. 11-13. In Fig. 11, one can observe that at low blink rates, other rules pertaining to target distance (rules 2-3) dominate the determination of robot speed, but at higher blink rates, the interpretation of user intent becomes more influential to increase speed (rule 1). In Fig. 12, the target distance is seen to have a positive correlation with speed, and speed increases up to the hardware-limited threshold. The adjustment of heading with obstacle proximity is shown in Fig. 13, where avoidance behavior is more extreme for closer obstacles (rules 13-15). All these behaviors are as intended and are consistent with the results of simulation.

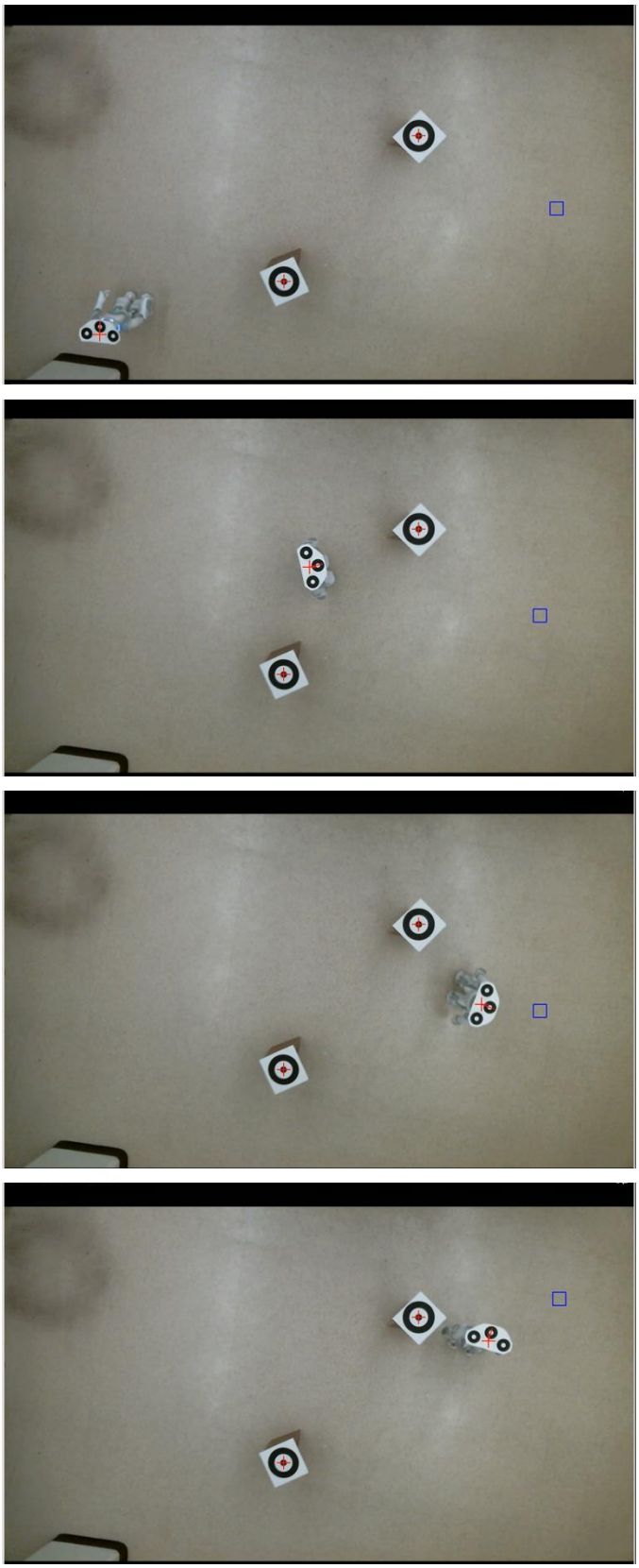


Figure 9. Experiment setup (from top to bottom): robot in starting position (lower left of workspace), robot navigating between obstacles, robot approaching first target point, robot approaching the second target point.

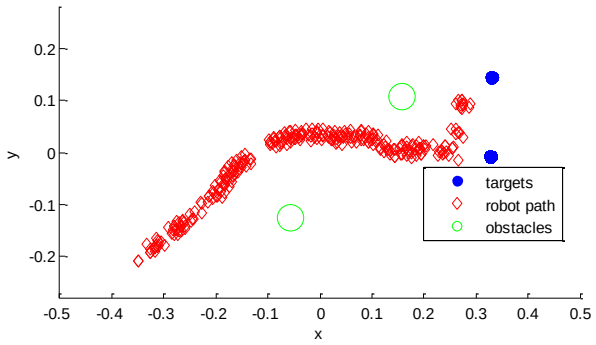


Figure 10. Robot trajectory from the experiment of Fig. 9.

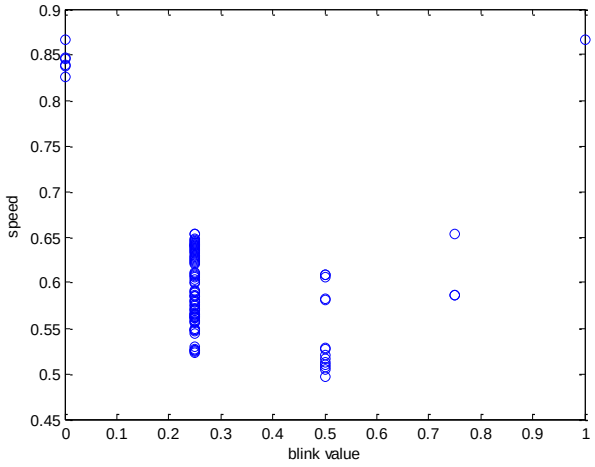


Figure 11. Modulation of speed based on blink rate demonstrates effective interpretation of user intent in the experiment of Fig. 9.

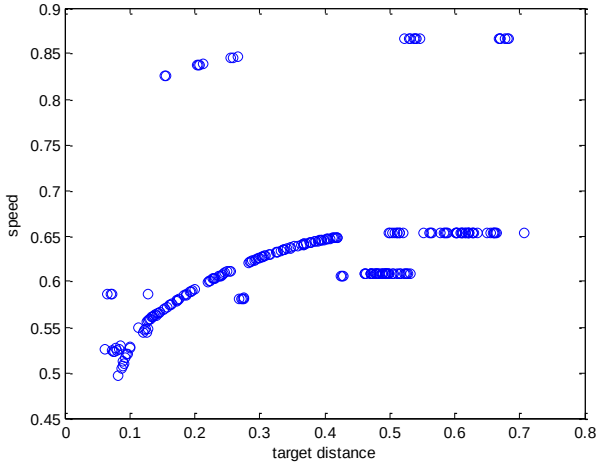


Figure 12. Modulation of speed based on target distance demonstrates effective target seeking in the experiment of Fig. 9.

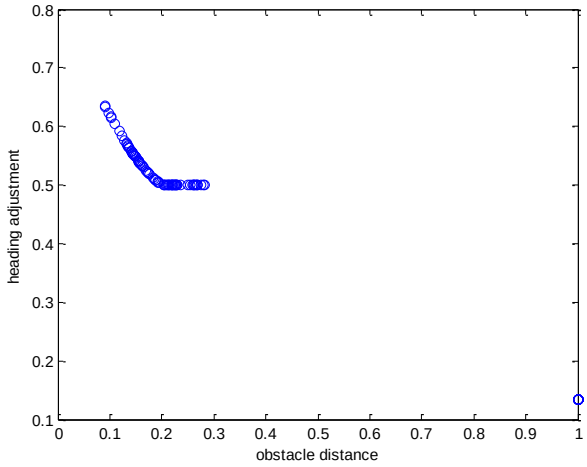


Figure 13. Modulation of heading adjustment based on obstacle distance demonstrates effective obstacle avoidance in the experiment of Fig. 9.

IV. CONCLUSIONS

In this paper, a technique for gaze-based guidance of personal assistance robots has been illustrated via simulation and experiments. Fuzzy logic allows the robot to simultaneously manage multiple behaviors, practicing energy conservation when appropriate but pursuing the target when human intent to do so is clear. Combined use of the eye gaze point and blinking data is a pivotal feature of the fuzzy logic controller. Basic obstacle avoidance is demonstrated as an integrated behavior within this controller. Additionally, the fuzzy controller was successfully used to control a real-time robot using actual gaze data acquired from human users using an eye tracking system.

The results presented in this paper suggest promise for additional future work, which could focus on incorporating the controller into a system that actively detects the robot's location and obstacles without the need for special markers. The controller should also be tuned for improved performance, and some of its more basic rules may be replaced by a more sophisticated steering and obstacle avoidance rule set based on recent research in inference modeling [21][22]. Performance comparison with other MIMO control approaches will then be appropriate. More advanced work will focus on detailed implementation for a broader variety of personal assistance tasks (e.g., object pick-and-place, operating on a static object) in a true 3D environment.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grants No. 1264504 and 1414299.

REFERENCES

[1] C. Nelson, "Fuzzy Logic Control for Gaze-Guided Personal Assistance Robots," Proc. Third International Conference on Intelligent Systems and Applications, Seville, Spain, June 22-26, 2014, pp. 25-28, 2014.

[2] J. Woodfill, R. Zabih, and O. Khatib, "Real-time motion vision for robot control in unstructured environments," Proc. ASCE Robotics for Challenging Environments, pp. 10-18, 1994.

[3] P. K. Allen, A. Timcenko, B. Yoshimi, and P. Michelman, "Automated tracking and grasping of a moving object with a robotic hand-eye system," IEEE Transactions on Robotics and Automation, vol. 9, no. 2, pp. 152-165, 1993.

[4] T.-S. Jin, J.-M. Lee, and H. Hashimoto, "Position control of mobile robot for human-following in intelligent space with distributed sensors," International Journal of Control, Automation, and Systems, vol. 4, no. 2, pp. 204-216, 2006.

[5] J. Satake and J. Miura, "Robust stereo-based person detection and tracking for a person following robot," Proc. IEEE International Conference on Robotics and Automation 2009, Workshop on People Detection and Tracking, Kobe, Japan, May 2009.

[6] J. Han, S. Han, and J. Lee, "The tracking of a moving object by a mobile robot following the object's sound," J. Intell. Robot. Syst., vol. 71, pp. 31-42, 2013.

[7] C.-H. Chen, C. Cheng, D. Page, A. Koschan, and M. Abidi, "A moving object tracked by a mobile robot with real-time obstacles avoidance capacity," Proc. of the 18th International Conference on Pattern Recognition (ICPR 2006), 4 p.

[8] M. Mucientes and J. Casillas, "Learning fuzzy robot controllers to follow a mobile object," International Conference on Machine Intelligence, Tozeur, Tunisia, Nov. 5-7, 2005, pp. 566-573.

[9] M. Abdellatif, "Color-based object tracking and following for mobile service robots," International Journal of Innovative Research in Science, Engineering and Technology, vol. 2, no. 11, pp. 5921-5928, 2013.

[10] R.-E. Precup, S. Preitl, M.-B. Rădac, E. M. Petriu, C.-A. Dragoş, and J. K. Tar, "Experiment-based teaching in advanced control engineering," IEEE Transactions on Education, vol. 54, no. 3, pp. 345-355, 2011.

[11] X. Zhang, S. Li, J. Zhang, and H. Williams, "Gaze Contingent Control for a Robotic Laparoscope Holder," J. Med. Devices, vol. 7, no. 2, pp. 020915.1-020915.2, 2013.

[12] D. P. Noonan, G. P. Mylonas, A. Darzi, and G.-Z. Yang, "Gaze Contingent Articulated Robot Control for Robot Assisted Minimally Invasive Surgery," Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, Nice, France, Sept. 22-26, 2008, pp. 1186-1191.

[13] R. Barea, L. Boquete, L. M. Bergasa, E. López, and M. Mazo, "Electro-oculographic guidance of a wheelchair using eye movements codification," The International Journal of Robotics Research, vol. 22, no. 7-8, pp. 641-652, Jul. 2003.

[14] C. S. Lin, C. W. Ho, W. C. Chen, C. C. Chiu, and M. S. Yeh, "Powered wheelchair controlled by eye-tracking system," Optica Applica, vol. 26, no. 2-3, pp. 401-412, 2006.

[15] P. S. Gajwani and S. A. Chhabria, "Eye motion tracking for wheelchair control," International Journal of Information Technology and Knowledge Management, vol. 2, no. 2, pp. 185-187, 2010.

[16] M. I. Shahzad and S. Mehmood, "Control of articulated robot arm by eye tracking," Master's Thesis no. MCS-2010-33, Blekinge Institute of Technology, Sep. 2010.

[17] R. Atienza and A. Zelinsky, "Intuitive human-robot interaction through active 3D gaze tracking," Robotics Research: The 11th International Symposium, pp. 172-181, 2003.

[18] A. Leonel, F. B. de Lima Neto, S. C. Oliveira, and H. S. B. Filho, "An Intelligent Human-Machine Interface Based on Eye Tracking to Afford Written Communication of Locked-In Syndrome Patients," Learning and Nonlinear Models, vol. 9, pp. 249-255, 2011.

[19] P. Hájek, Metamathematics of Fuzzy Logic, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[20] A. Glenstrup, Eye Controlled Media: Present and Future. Bachelor's Thesis in Information Psychology at the Laboratory of Psychology, University of Copenhagen, DK-2100, 1995.

[21] E. Martínez-Martín, M. T. Escrig, and A. P. del Pobil, "Naming qualitative models based on intervals: A general framework," International Journal of Artificial Intelligence, vol. 11, no. A13, pp. 74-92, 2013.

[22] X.-Z. Wang, J.-H. Zhai, and S.-X. Lu, "Induction of multiple fuzzy decision trees based on rough set technique," Information Sciences, vol. 178, no. 16, pp. 3188-3202, 2008.

ℜeALIS2.1: The Implementation of Generalized Intensional Truth Evaluation and Expositive Speech Acts in On-Going Discourse

Gábor Alberti and László Nőthig
University of Pécs, Department of Linguistics
ℜeALIS Theoretical, Computational and Cognitive Research Team, Pécs, Hungary
emails: {alberti.gabor@pte.hu, nothig.laszlo@gmail.com}

***Abstract*—Our software application ℜeALIS2.1 (just like ℜeALIS1.1) is primarily intended to supply linguists with a highly intelligent device to build fragments of languages. On the basis of the fragments, (non-linguist) experts can elaborate a peculiarly "multiplied" database that offers, besides the model of the external world, hundreds of its (appropriately labeled) alternatives. According to the ℜeALIS theoretical framework that we use (ℜeALIS: *Reciprocal and Lifelong Interpretation System*), these alternative models can all be linked to simulated human agents (addressers and addressees of possible discourses), who are represented in the world model as conglomerates of their pieces of knowledge, beliefs, desires, and intentions. Finally, (further) users can select lexical items to build sentences, the truth-conditional interpretation of which the program can calculate on the basis of the actual version of the above-sketched "multiplied world model". It performs this after checking whether the given sentences can serve as felicitous expositive speech acts in realistic on-going discourses. Our software application serves not only the theoretical purpose of testing ℜeALIS as a Discourse-Representation-Theory-based "pragmalinguistics" approach (by implementing it), but it also serves the practical purpose of collecting and systematizing data in the peculiar structure that ℜeALIS offers. It is a crucial feature of ℜeALIS that it is intended to truly capture human intelligence (more precisely, such sapiens-specific components of long-term memory as episodic memory, with its space-time coordinates, and semantic memory, containing context-free knowledge).**

Keywords—dynamic discourse semantics; possible worlds; truth-conditional interpretation; speech acts; presupposition.

I. INTRODUCTION

We are working on the implementation of a pragma-linguistics theory, ℜeALIS, intended to truly capture human intelligence by means of a peculiarly multiplied world model [1] [2] [3]. We consider the implementation of this "intelligent" structure [4] [5] [6] our primary innovation.

The point of departure to our research work is a linguistics theory whose representatives, in the course of describing human language(s) and explaining their structure and functioning, have been led to a conclusion that may seem to be quite strange at first glance: revealing the "internal" secrets of language (including non-pragmatic and non-cognitive phenomena as well) depends on the substantial capturing of an external factor. This factor is information states of human minds in communication, changing from second to second [7] [8]. It is this factor that must be modeled in a way that we can account for the facts that we are interlocutors reciprocally "reflecting" each other's minds [9] [10] [11] and that our momentary information states contain even pieces of information obtained decades earlier. In essence, it is the human mind itself that is to be modeled according to special aspects and requirements [12] [13].

Is this a huge cost for the treatment of internal questions of language?

Our answer is 'no' to this question, because it is possible to elaborate a sufficiently simple plausible mathematical model [3] (see Section III). The promising benefit, however, opens up new prospects in two fields, in which we intend to continue to conduct research, besides linguistic phenomena in a narrow sense of the term. One field is basing the innumerable kinds of computational processing of human language upon data arranged according to human intelligence or the "model of minds in communication" [11] [14] [15]. The other field has to do with the scientific description of mental disorders: it is via inspecting the impaired mind, on the one hand, that we can approach to understanding the driving forces for language, and, on the other hand, it is via studying language that the decisive features of autism or schizophrenia, for instance, can be captured [16]. In the neuropsychiatric field of our research, we explain what causes information loss and deficiency in these conditions [17].

After sketching this broad picture, we restrict ourselves, in what follows, to dealing with the pragmalinguistics [5], mathematical [3] and "technical" [18] [19] [20] apparatus of ℜeALIS, which makes its implementation immediately possible.

Let us now overview the structure of the paper. Section II sketches the current version of ℜeALIS, primarily its radical ontological innovation relative to Discourse Representation Theory (DRT [21] [22]), which underlies it. Then the decisive elements of the mathematical definition of ℜeALIS are presented, in Section III. Section IV is devoted to the demonstration of the new results in pragmatics in the ℜeALIS framework, which have strengthened our earlier guideline. The point is that in the case of an utterance, it is to be checked whether the speaker, the hearer and the given situation are suitable for serving as the addresser, the addressee and the context of the linguistically defined speech act [23] [24], which simply requires a truth-conditional investigation [25] primarily into the addresser's mind's certain "worldlets". The task boils down to get to the worldlets in which certain polarity values must then be checked. Then our software application ℜeALIS1.1 is

demonstrated in Section V through discussing its different kinds of potential users and its main use cases for the users we call internal users and for those we call external users. Section VI demonstrates the analysis of some linguistic examples with the purpose of elucidating our ambition to capture the highest possible level of human intelligence coded in language. It is presented how our generalized truth evaluation can be applied to such complicated linguistic phenomena as tense, aspect, subjectivity, deixis, among others. Finally, Section VII presents the additional services of ℜeALIS2.1 as compared to ℜeALIS1.1 and an SDRT-based (Segmented DRT [22])) experimental software application called RUDI [26]. We point out that ℜeALIS2.1 can be regarded as a model of the two parts of long-term memory—episodic and semantic memory—and this enables us to derive a potentially infinite number of senses for words from finite lexical resources.

II. THE CURRENT VERSION OF ℜEALIS AND THE "STATE OF THE ART"

ℜeALIS is based on Discourse Representation Theory, often referred to as DRT [27] [28]; it can thus be introduced as belonging to the family of representational dynamic discourse semantics. Its complete (forty-page-long mathematical) definition is available at [2]; the relevant details will be given in Section III. It is intended in ℜeALIS to reconcile the formal exactness of generative syntaxes [29] [30] [31] (and their adaptations to Hungarian [32] [33] [34]) and the dynamic approach of optimality theories [35] and the aforementioned (S)DRT with basically Austinian [36] speech-act theories [37] [38], bearing in mind the holistic stance of cognitive linguists [39] [40].

In the post-Montagovian world [25] of formal semantics, DRT—which has offered a revolutionary logics-based solution to the resolution problem of ("donkey") anaphora and attractive visual representations for discourse meaning—is often criticized from "inside" as well as from "outside", considerably weakening its legitimacy. The internal criticism comes from the world of the dynamic model-theoretic semantics, from the Amsterdam School [41], and pertains to the (mathematically unquestionable) eliminability of exactly this attractive visual representation, insisting on "Montague's heritage" [25]. The external criticism comes from experts of philosophy /pragmatics [42] and representatives of the Proof-Theoretic School [43], among others [44]; they all point at the dubious status and construction of *possible worlds* (among others).

Pollard [44], for instance, is led to the following conclusion pertaining to the mainstream Kripke/Montague-inspired possible-worlds semantics: "the idea of taking worlds as a primitive of semantic theory is a serious misstep." He calls it [44] "a framework known to have dubious foundations."

Even the seminal book of teaching Montague Grammar [25] admits these "dubious foundations" in the course of discussing the problem of necessity and possibility: "Would this be an enlightening way of analyzing the semantics of necessity [e.g., *Alfred must be a bachelor*] and possibility [e.g., *Alfred may be a bachelor*]? Many philosophers of language have unequivocally answered "no" to this question; they have contended that since "possible worlds" are surely vague and ill-understood entities..., it cannot help to explain one mysterious semantic concept (necessity) in terms of an even more mysterious one (possible worlds)."

The same is still "reported" in 2014.

Judge [45], for instance, who works in the standard, Kratzerian [46] [47] framework of modality (based on the Kripke/Montague-inspired possible-worlds semantics), "admits" that "describing the semantics of uncertainty is problematic – particularly for semantic theories that are reliant on truth-conditional definitions of meaning;" and she designates the pertinent relationship between formal semantics and pragmatics as follows: "...ideally a linguistic theory will account for how natural language works in real conversational contexts, and not be restricted to only accounting for logical output, (not least because extricating the core/logical meaning of a linguistic expression from the contributions of context is highly problematic). Indeed, modality is an area of semantics where understanding the systematic interactions of context and underlying form is particularly pertinent." Note that Judge's evaluation even on her own solution proposed in [45] is definitely low-key: "The proposal of the certainty set is intended as an experiment, rather than a full-blown, conclusive solution to the puzzles of modal expressions. By refashioning the knowledge set as a certainty set some interesting patterns and solutions are suggested. However, problems remain particularly with characterising degrees of modality, the epistemic modality/ evidential" distinction, [among others]..."

Marsali [42], whose approach is philosophical /pragmatic, "...refuses to adopt the semantic account of EMM [epistemic modality markers, such as *maybe*, *probably*, *certainly*, *definitely*] on the ground of ... [the reason that] it is not clear how EMM should be interpreted, and countless incompatible semantic accounts of EMM have been presented in the philosophical literature ... But it is implausible to contend that EMM fix the truth conditions of [say, a] statement like *Certainly it is raining in England*, if there is no agreement on what are the truth conditions of [a statement like this]."

It is also worth mentioning on the basis of [7] that "221 *may* be a good choice" is as reasonable a reaction to the proposal "We need a prime number greater than 200" as the reaction "211 *may* be a good choice," in contrast to the unreasonable reaction "300 *may* be a good choice." The problem (for possible-worlds semantics) is that 211 is a prime number, indeed, while $221=13\cdot17$. There is no possible world, thus, in which 221 is a prime number (or, in an absurd system of possible worlds, even 300 can qualify as a prime number).

We claim that ℜeALIS—while considerably relying on the representationalism of DRT in the course of solving a wide range of linguistic problems in order to maximally exploit and develop the excellent facilities provided by this representationalism—offers exactly the radical ontological innovation that has to do with the elimination of the above-

mentioned two dubious levels of representation, discourse representations and possible worlds, referred to as I and III in Fig. 1.

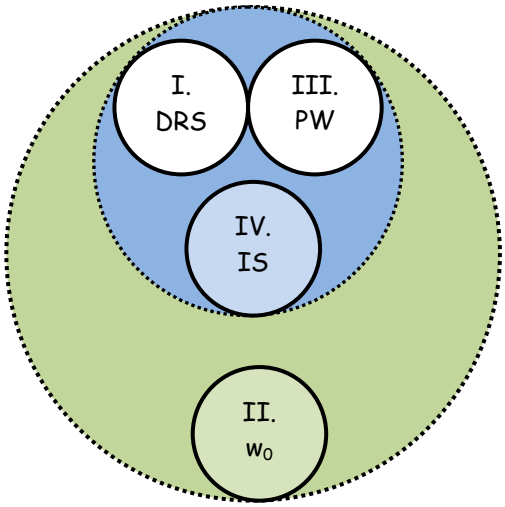


Figure 1. Components / levels of representation in DRT: I-IV; and their re-arranged ontology in ℜeALIS:
I. DRS: the semantic representation of sentences constituting coherent texts
II. Model of the external world (for extensional interpretation)
III. Possible worlds (for intensional interpretation)
IV. Interlocutors' information states

ℜeALIS embeds representational levels I and III—more exactly, their relevant content—in the representation of information states (IV), relying on the approach that, as interlocutors obtain information through discourses, their information states are worth regarding as gigantic, lifelong, DRSs. An information state has a double nature: it functions as a "representation" in the above regard while it is used as "what is to be represented" in the interpretation of, say, the intensional sentence types shown in (2b-d): it also depends on different persons' information states whether these sentences are true, in contrast to sentence (2a), the truth value of which only depends on facts of the external world. Note in passing about the aforementioned "double nature of information states" that modern set theory exactly rests upon a similar idea: Sets and their elements must not be mixed up; this does not mean, however, that a set could not serve as an element of another set.

a. "Ben is a linguist."
b. "Sue knows that/if [Ben is a linguist]"
c. "Joe guesses that Sue definitely wants to convince him to take it for granted that [Ben is a linguist]."

Figure 2. Sentences to be interpreted in different world(let)s.

We are now going to illustrate the descriptive and explanatory power of ℜeALIS by sketching the interpretation of sentence (3a), featuring *realize*, which is a factive verb (NB: similar analyses of ours are available in [8] and [7]). Hence, it is a precondition of interpreting the sentence as true (or rather, as "well-formed") that the Evening Star must coincide with the Morning Star in (the model of) the external world. This means that the entity referred to as the Evening Star by the given astronomer must be the same entity he refers to as the Morning Star. In the approach of ℜeALIS, this relation is captured formally as demonstrated in (3b): the internal entity $r_{EveningStar}$ must be anchored to the same external entity as the internal entity $r_{MorningStar}$.

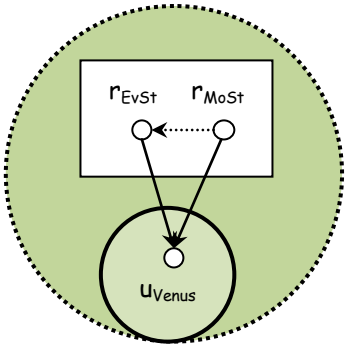


Figure 3. The interpretation of *realize* and the *Venus*-problem
a. "An ancient astronomer realized that the Evening Star is the same as the Morning Star."
b. $\alpha(r_{EvSt})$ is-the-same-as $\alpha(r_{MoSt})$ (since u_{Venus} is-the-same-as u_{Venus})
c. It does not hold that $r_{EveningStar}$ is-the-same-as $r_{MorningStar}$ at τ in the astronomer's worldlet of astronomic hypotheses
d. It holds that $r_{EveningStar}$ is-the-same-as $r_{MorningStar}$ at τ', which is a later point of time in the astronomer's worldlet of astronomic hypotheses

The astronomer himself is not (necessarily) aware of the co-anchoring of the two internal entities at his disposal (in his appropriate worldlet); but the fact of co-anchoring is an external requirement due to the factive character of the verb. Two further requirements to be satisfied in order for sentence (3a) to qualify as true concern two information states of the astronomer at different points of time, independently of the external world: what is to be checked is whether there is a "same-as" relation between the internal entity $r_{EveningStar}$ and the internal entity $r_{MorningStar}$ in the one information state (3d) while they do not stand in the "same-as" relation in the other one (3c)

All in all, three competing world(let) models need to be considered simultaneously ("prism effect"), and three entities—an external one and two internal ones—need to be inspected. As the three models are all parts of the one complete model of the history of the external world and all internal reflections associated with it (see Fig. 2), in this matrix model (3b-d) can all be checked.

It must be noted that the analysis relies on the same facilities available in the cognitive linguistics framework; see, for instance, the paper by Pelyvás [39], who follows Langacker's approach to nominal grounding [48]. The most important tenet of this view is that all nominals are grounded in the "reality" of the Idealized Conceptual Model(s) evoked in the discourse, which is relative to speaker and hearer, rather than directly in objective reality. From the point of view of linguistic analysis the reality that

we could call "objective" (i.e., independent of speakers' and hearers' beliefs) is only of marginal importance.

At this point we call the reader's attention to the obvious fact that our treatment of replacing, in the course of truth-conditional interpretation, a set of possible worlds with the finite (and typically very small) "worldlet" containing the information shared by the given possible worlds opens up new prospects in (the practice of) implementation. With gigantic sets of gigantic possible worlds got rid of, there is already no obstacle to capturing the pragmatic complexity that is claimed to be associated with even simple assertions, which "serve the aim of communicating, not merely pieces of information, but also the speaker's attitude of certainty or uncertainty about them," [38] chiefly due to what are called the ATMM-categories: Aspect [49], Tense, Mood, and (different kinds of) Modality [38] (see also [40], Section 4 in [24], and Section IV-D in the present paper).

We conclude this section by telling some words on the "state of the art". At the moment, we only have world models and alternative-worldlet-set models filled up with small sets of data. Our sophisticated Hungarian lexicon also consists of not more than a few hundreds of words [51] [52] [53], and our English lexicon is even smaller; furthermore, the involvement of these lexical items in parsing [54] also requires highly theory-specific morphological [55], syntactic [57] [56] [58], and semantic [1] [58] [60] tools. Thus, we are in an experimental phase with our software applications. The problem is that it would require very much time, effort, and, hence, money, to elaborate realistic and useful world(let) models. We are also in need of native speakers of English who are willing to help elaborating such sophisticated linguistic descriptions as the Hungarian ones presented in Section IV-D, for instance.

It would be worth elaborating all these costly components if, and only if, such customers appeared in the market who are willing to make our team register the data they work with and the environment they work in according to the peculiar system that ℜeALIS offers. We mean data that someone *actually* works with. The primary aim with this paper is to find such customers. Not only for obtaining financial support for our team of theoretical and computational linguists, most importantly for obtaining realistic conglomerates of data worth working up in the ways demonstrated in this paper.

We intend to convince the reader (and our potential customers) that not only professional spies, intrigants and mind readers are in need of maniulating data registered in multiplied wordlet structures but also detectives, lawyers and judges are, as well as managers and secretaries, psychiatrists and politicians, and practically everyone. It is no coincidence that every human language is well equipped with such communication tools and techniques as those presented in Section IV.D and in certain subsections of Section VI. It may turn out to be important to anyone to be aware of such complex epistemic patterns (concerning human agents A_1, A_2, A_3,... and potential facts ψ_1, ψ_2, ψ_3,...) as the situations sketched in the following paragraphs.

A_1 and A_2 both know ψ with(out) knowing this about each other.

A_1, who knows ψ, wants A_2 to think that he does not know ψ or that he thinks so that ψ is probably false.

A_1 wants to get known from A_2 whether ψ is true or false, but he does not want her to notice this intention.

A_1, who is telling A_2 that ψ is true, is almost sure that A_2 is convinced that he does know whether ψ is true or false.

A_1 knows that A_2 is aware of the fact that ψ is true but he pretends as if he did not know that.

A_1 did not know whether ψ_1 or ψ_2 is true out of two incompatible statements but he had to make a decision. He knew that ψ_1 is true according to A_1, A_2, and A_3, while ψ_2 is true according to A'_1, A'_2, and A'_3. Now he thinks so that ψ_1 is probably true while ψ_2 is probably false. He has made this decision on the basis of the following facts and assumptions: A_1, A_2, and A_3 have proved reliable people in similar cases in which decisions had to be made, in contrast to A'_1, A'_2, and A'_3. Moreover, A_1 suspects that A'_1, A'_2, and A'_3, who have close contact with each other, are interested in his believing in ψ_2, whereas A_1, A_2, and A_3 are likely to have never met each other.

III. The Decisive Elements of the Definition of ℜeALIS

The relevant parts of the mathematical definition of ℜeALIS (whose 40 page long complete version is available in [2]) are summarized here. As interpreters' mind representation is part of the world model, the definition of this model $\mathfrak{R} = \langle U, \mathbf{W_0}, W\rangle$ is a complex structure where

- U is a countably infinite set: the universe;
- $W_0 = \langle U_0, \mathbf{T}, \mathbf{S}, \mathbf{I}, \mathbf{D}, \Omega, \mathbf{A}\rangle$: the external world;
- W is a partial function from I×T where W[i,t] is a quintuple $\langle U[i], \sigma[i,t]^{\Pi}, \alpha[i,t]^{\Psi}, \lambda[i,t]^{\Lambda}, \kappa[i,t]^{K}\rangle$: the internal-world function.

The external world consists of the following components:

- U_0 is the external universe ($U_0 \subset U$), whose elements are called entities;
- $\mathbf{T} = \langle T, \Theta\rangle$ is a structured set of temporal intervals;
- $\mathbf{S} = \langle S, \Xi\rangle$ is a structured set of spatial entities;
- $\mathbf{I} = \langle I, Y\rangle$ is a structured set of interpreters;
- $\mathbf{D} = \langle D, \Delta\rangle$ is a structured set of linguistic signs (practically morph-like entities and bigger chunks of discourses performed),
- where $T \subset U_0$, $S \subset U_0$, $I \subset U_0$, $D \subset U_0$,
- $\Omega \subset T \times U_0^*$ is the set of core relations (with time intervals as the first argument of all core relations),
- **A** is the information structure of the external world (which is nothing else but relation structure Ω reformulated as a *standard simple information structure*, as is defined in [61]; its basic elements are called the *infons* of the external world.

The above mentioned *internal-world function* W is defined as follows:

- The relation structure W[i,t] is called the internal world (or information state) of interpreter i at moment t;

- $U[i] \subset U$ is an infinite set: interpreter i's internal universe (or the set of i's referents, or internal entities); U[i'] and U[i"] are disjoint sets if i' and i" are two different interpreters;
- what changes during an interpreter i's lifespan is not her referent set U[i] but only the four relations among the (peg-like [62] [8]) referents, given below, which are called i's internal functions:
- $\sigma[i,t]^{\Pi} : \Pi \times U[i] \rightarrow U[i]$ is a partial function: the eventuality function (where Π is a complex label characterizing argument types of predicates),
- $\alpha[i,t]^{\Psi} : \Psi \times U[i] \rightarrow U[i] \cup U_0$ is another partial function: the anchoring function (α practically *identifies* referents, and Ψ contains complex labels referring to the legitimizing grammatical factors);
- $\lambda[i,t]^{\Lambda} : \Lambda \times U[i] \rightarrow U[i]$ is a third partial function: the level function (where elements of Λ are called level labels); the level function is intended to capture the "box hierarchy" among referents in complex Kampian DRS boxes [21] enriched with some rhetorical hierarchy in the style of SDRT [22],
- $\kappa[i,t]^{K} : K \rightarrow U[i]$ is also a partial function: the cursor, which points to certain temporary reference points prominently relevant to the interpreter such as "Now", "Here", "Ego", "Then", "There", "You" [38];
- The temporary states of these four internal functions above an interpreter's internal universe serve as her "agent model", or mind representation, in the process of (static and dynamic) interpretation.

Suppose the information structure **A** of the external world (defined above as a part of model $\mathfrak{R} = \langle U, \mathbf{W_0}, W \rangle$) contains the following infon: $\iota = \langle$perceive, t, i, j, d, s$\rangle$, where i and j are interpreters, t is a point of time, s is a spatial entity, d is a discourse (chunk), and 'perceive' is a distinguished core relation (i.e., an element of Ω). The interpretation of this "perceived" discourse d can be defined in our model relative to an external world $\mathbf{W_0}$ and internal world W[i,t].

The dynamic interpretation of discourse d is essentially a mapping from W[i,t], which is a temporary information state of interpreter i, to another (potential) information state of the same interpreter that is an *extension* of W[i,t]; which practically means that the above mentioned four *internal functions* (σ, α, λ, κ) are to be developed monotonically by *simultaneous recursion*, expressing the addition of the information stored by discourse d to that stored in W[i,t].

The new value of eventuality function σ chiefly depends on the *lexical items* retrieved from the interpreter's internal mental lexicon as a result of the perception and recognition of the words / morphemes of the interpreter's mother tongue in discourse d. This process of the identification of lexical items can be regarded as the first phase of the dynamic interpretation of (a sentence of) d. In our ℜeALIS framework, extending function σ corresponds to the process of accumulating DRS condition rows containing referents that are all—still—regarded as different from each other.

It will be the next phase of dynamic interpretation to *anchor* these referents to each other (by function α) on the basis of different grammatical relations that can be established due to the recognized *order* of morphs / words in discourse d and the *case*, *agreement* and other markers it contains. In our approach, two referents will never have been *identified* (or deleted), they will only be anchored to each other; but this anchoring essentially corresponds to the identification of referents in DRSs.

The third phase in this simplified description of the process of dynamic interpretation concerns the third internal function, λ, the level function. This function is responsible for the expression of intra- and inter-sentential scope hierarchy [63] [64] / information structure [32] [33] [34] / rhetorical structure [22], including the embedding of sentences, one after the other, in the currently given information state by means of rhetorical relations essentially in the way suggested in SDRT.

It is to be mentioned that the information-state changing dynamic interpretation and the truth-value calculating *static interpretation* are mutually based upon each other. On the one hand, static interpretation operates on the *representation* of sentences (of discourses) that is nothing else but the output result of dynamic interpretation. On the other hand, however, the above discussed phases of dynamic interpretation (and chiefly the third phase) include subprocesses requiring static interpretation: certain *presuppositions* are to be verified [21] [38].

The interpreter's fourth internal function, cursor κ, plays certain roles during the whole process of dynamic interpretation. *Aspect*, for instance, can be captured in our approach as the resetting or retaining of the *temporal* cursor value as a result of the interpretation of a sentence ($\rightarrow$ non-progressive / progressive aspect, respectively). It can be said in general that the input cursor values have a considerable effect on the embedding of the "new information" carried by a sentence in the interpreter's current information state and then this embedding will affect the output cursor values [65].

Dynamic interpretation in a ℜeALIS model $\mathfrak{R}=\langle U, \mathbf{W_0}, W \rangle$, thus, is a partial function Dyn that maps a (potential) information state W° to a discourse d and an information state W[i,t] (of an interpreter i):

- $\mathrm{Dyn(d)} : \langle \mathfrak{R}, W[i,t] \rangle \mapsto \langle W^{\circ}, \underline{e^{\circ}}, \mathbf{U^{\circ}} \rangle$,
- where $\mathbf{U^{\circ}}$, shown up in the output triple, is the cost of the given dynamic interpretation (coming from presuppositions legitimized by *accommodation* instead of *verification*), and $\underline{e^{\circ}}$ is the eventuality that the output cursor points to (this is the eventuality to be regarded as representing the content of discourse d). Function Dyn(d) is *partial*: where there is no output value, the discourse is claimed to be ill-formed in the given context. Due to the application of cost, ill-formedness is practically a gradual category in ℜeALIS.

The static interpretation of a discourse d is nothing else but the static interpretation of the eventuality referent that represents it. The recursive definition of static interpretation

is finally based upon anchoring internal entities of interpreters to external entities in the external universe, and advances from smaller units of (the sentences of) the discourse towards more complex units.

The "prism effect", mentioned in Section II, is worth to be given a separate definition in the system of ℜeALIS, as follows, because our linguistic illustrations are practically based on this single formula.

A clause performed in a context conveys an infon that belongs to an *intensional profile*, which is an element of the set defined below: the power set of the set of finite sequences of a particularly specialized set of the above-defined level labels. The clause is to be interpreted against the (possible-world-like but finite) components of this intensional profile in order to obtain its truth conditions and other semantic and/or pragmatic well-formedness conditions in the given context.

$$\mathscr{P}((\mathrm{M}\times\mathscr{P}(\mathrm{I})\times\mathrm{P}\times\mathrm{T}\times\mathscr{P}(\{+,-,0\}))^*) = \mathscr{P}(\Lambda^*) \qquad (1)$$

The theoretically highly important mathematical exactness, which this formula is intended to suggest, provides a simple, straightforward, uniform, well motivated, and "user-friendly" approach to reach the ultimate aim of pragmalinguistics [66] [67]: to account for the use of the semantic content of the sentences performed in a certain context (see also [45]).

Let us start the elaboration of the details with set M in formula (1): it is the set of modal labels that say whether an infon serves to someone as some kind of belief (BEL), or desire (DES), or intension (INT), or anything else ([5]; see also [38]). Mann [68] establishes that "[p]erhaps the most important distinction for modeling the intentions that accompany language use is a contrast between intended actions and intended effects. Intended effects typically are states of affairs that the intender desires or prefers, while intended actions typically involve some identifiable process within the capacities of the actor(s)." On the basis of this, 'intended effects' are called desires (DES) in ℜeALIS.

Set I provides degrees for expressing the intensity of the given modality, from "maximum" (MAX or M) through "great" (gr) up to "some" (sm). Associated with the modality BEL, for instance, this scale ranges from sure knowledge to weak conjecture. There must be "uncertain" degrees between "known" and "unknown" [69]. The *muss/soll/will* triplet of German epistemic modal auxiliary verbs can be regarded as evidence for the existence of at least three non-maximal degrees [38]. It requires much future research to decide how many degrees have a linguistic relevance in the certainty-uncertainty continuum [42]. The power set $\mathscr{P}(\mathrm{I})$ of I is used in formula (1), because certain modal words may be associated with more than one degree of intensity of a given modality.

Set R is responsible for referring to the host of the given infon, who can primarily be the speaker (MY) or the hearer (YR: 'your'). That is, the possible-world-like (but finite) basis of interpretation (1), called a "worldlet" in ℜeALIS [13], can be the conglomerate of "my faint conjectures" or "your strong desires", and so on.

Set T adds "temporal stamps" to worldlets, expressing in which period it holds that a given infon belongs to a given worldlet in someone's mind (to the one, for instance, that stores someone's faint conjectures).

Worldlets are also assigned polarity values, which are members of the eight-element powerset $\mathscr{P}(\{+,-,0\})$ of the two traditional polarity values "true" (+) and "false" (–) and a not so accustomed value "non-specified" (see the category "unknown" in [69]). The crucial importance of the fact that the traditional two-element set of truth values has been extended to an eight-element one is illustrated by the difference between the interpretation of (4a) and (4b). In the latter case what is certain is that the given infon ("Ben is a linguist") is not assigned 0 in Sue's mind; but sentence (4b) does not reveal whether it is true or it is false that Ben is a linguist.

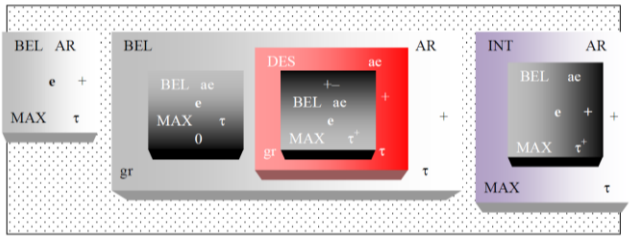
a. "Sue knows that [Ben is a linguist]"
$\langle \mathrm{BEL},\mathrm{MAX},r_{\mathrm{Sue}},\tau,+\rangle$
b. "Sue knows if [Ben is a linguist]"
$\langle \mathrm{BEL},\mathrm{MAX},r_{\mathrm{Sue}},\tau,+-\rangle$
c. "Joe guesses that Sue definitely wants to convince him to take it for granted that [Ben is a linguist]."
$\langle\langle \mathrm{BEL},\mathrm{sm},r_{\mathrm{Joe}},\tau,+\rangle\langle \mathrm{INT},\mathrm{MAX},r_{\mathrm{Sue}},\tau',+\rangle\langle \mathrm{BEL},\mathrm{MAX},r_{\mathrm{Joe}},\tau'',+\rangle\rangle$

Figure 4. Sentences to be interpreted in different world(let)s.

The Kleene-star in formula (1) manifests the "reciprocal" character of ℜeALIS by offering, instead of quintuples of the above-discussed labels, finite series of such quintuples. The series shown in (4c), for instance, points to a special segment of Joe's mind: namely, to the worldlet containing Joe's hypotheses on Sue's intentions towards exactly on his would-be information state.

Finally, the power set symbol in the initial position of formula (1) requires some explanation. The point is that an infon (a piece of information) can be simultaneously associated with more series of worldlet labels (in the human mind). The reference to a "prism effect" in (1) expresses this viewpoint.

The content of the components in Fig. 5, for instance, applied to the Hungarian declarative sentence in (5a), is as follows, from left to right: "(5b) I, (the addresser: AR) know that Péter moved to Mari (I refrain from telling lies or bluffing). (5c) I think that you (the addressee: ae) do not know this. (5d) I think that you would like to be aware of this fact at a later point τ^+ in time (otherwise, I would not have uttered the sentence, since it is important for me to be relevant). (5e) (Being also cooperative) I intend to help you to acquire the infon in question." This analysis is based on the Gricean maxims of conversation [70]; further details are available in our following papers: [5] [3]. The visual representation is essentially a conglomerate of (S)DRT boxes, but, instead of parts of segmented logical formulas, it is immediately the referents (constants/variables) contained that are placed in the partially ordered boxes (in the form of Landmanian "pegs" [62] [8]), augmented with the aforementioned level labels.



a. Péter Marihoz költözött.
P. M.Ade move.Past.3Sg
'Péter moved to Mari's.'
b. ⟨B,M,AR,τ,+⟩
c. ⟨B,gr,AR,τ,+⟩⟨B,M,ae,τ,0⟩
d. ⟨B,gr,AR,τ,+⟩⟨D,M,ae,τ,+⟩⟨B,M,ae,τ^+,+–⟩
e. ⟨I,gr,AR,τ,+⟩⟨B,M,AR,τ^+,+⟩

Figure 5. The intensional profile of the Hungarian declarative sentence.

The conglomerate of the four components in (6b-e) and in Fig. 6 is the intensional profile of the simplest type of yes/no questions in Hungarian, as is proposed in [6] and [3]. Its specific content can be formulated in English as follows, compared to that of the declarative sentence. "1. Now it is me, the addresser, who does not know if Péter moved to Mari. 2. I think, however, that you know the truth. 3. I wish I also knew the truth. 4. (That is why I have started the conversation) I intend to help you to intend to help me to acquire the infon in question."

The circled fifth component presents the pragmatico-semantic contribution of the discourse particle *ugye* (see also the level label given in (6f)). Its contribution can be defined by simply adding a single component to the four-component representation of the yes/no question, which is responsible for expressing the speaker's bias towards the positive answer: "I consider it likely that Péter called Mari." Note in passing that it is no contradiction that the speaker conveys that (s)he is not absolutely sure that Péter moved to Mari (6b) but, at the same time, (s)he considers it quite likely (6f). In our approach, as was mentioned, different levels of knowledge (BEL/MAX vs. BEL/gr) can be considered, and can also be evaluated separately.

Our analysis of the discourse particle *vajon* [3] is based on the observations of Gärtner and Gyuris [71] and Schirm [72] that this special grammatical clue expresses "speculation", "hesitation", "uncertainty", "curiosity", and "reflection". Its meaning—or rather, its pragmatico-semantic contribution—can be revealed by comparing its intensional profile to the intensional profile given in (6b-e), which shows differences in two components out of the four.

The content of the components in Fig. 6.II can be paraphrased as follows: "1. I do not know if Péter moved to Mari's. 2. *I consider it likely* (*unfortunately*) that you do not know the truth either. 3. I wish I knew the truth. 4. (Why have I started the conversation, anyway?) I want you to know that I intend to acquire the given infon."

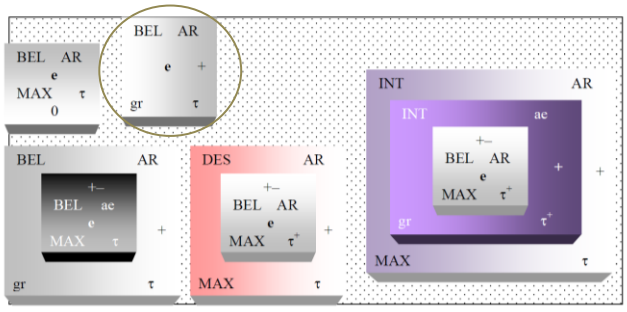


Figure 6.I ↑
a. Péter Marihoz költözött? a'. Péter *ugye* Marihoz költözött?
P. M.Ade move.Past.3Sg P. *ugye* M.Ade move.Past.3Sg
'Did Péter move to Mari's?' 'Péter moved to Mari's, did not he?'
b. ⟨B,M,AR,τ,0⟩
c. ⟨B,gr,AR,τ,+⟩⟨B,M,ae,τ,+–⟩
d. ⟨D,M,AR,τ,+⟩⟨B,M,AR,τ^+,+–⟩
e. ⟨I,M,AR,τ,+⟩⟨I,gr,ae,τ^+,+⟩⟨B,M,AR,τ^{++},+–⟩
f. *ugye*: ⟨B,gr,AR,τ,+⟩

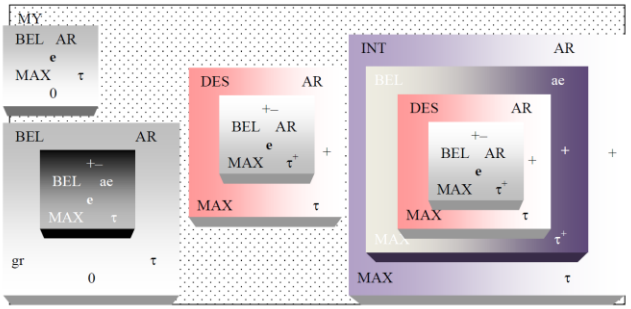


Figure 6.II ↑
g. Péter *vajon* Marihoz költözött?
P. *vajon* M.Ade move.Past.3Sg
cca. 'I would like to know whether Péter has moved to Mari's.'
Intensional profile of (6g): (6b) + (6d) + (6g') + (6g")
g'. ⟨B,gr,AR,τ,+⟩⟨B,M,ae,τ,0⟩
g". ⟨I,M,AR,τ,+⟩⟨B,M,ae,τ,+⟩ ⟨D,M,AR,τ,+⟩⟨B,M,AR,τ^+,+–⟩

Figure 6. The intensional profile of the basic Hungarian yes/no question type in 6.I and (6a), and that of its variants containing the discourse particles *ugye* (see I. and (6a'-f)) and *vajon* (see 6.II and (6g-g")).

Components 1 and 3 are common: the addresser, who does not know if a certain infon is true or false, longs for this knowledge. Components 2 and 4 are new. The addresser does not really hope that the addressee knows the answer; (s)he is only thinking aloud, with no immediate purpose. The only realistic purpose for him/her may be to make the addressee know that (s)he needs the answer.

It is worth addressing the division of labor between pragmatics and semantics. We can separate the at-issue meaning (coming from the original question) and the additional meaning (coming from the discourse particles). *Ugye* and *vajon*, discussed (briefly) above, bear the properties of conventional implicatures [73] [74], namely: semantic (lexical), independent (from at-issue content), secondary (supporting content—"fine-tuning"), not

backgrounded (not part of the common ground), not deniable, and invariably speaker-oriented. In both cases, the at-issue meaning denotes whether Péter moved to Mari's or not. As for *ugye*, the conventional implicature expresses bias toward the positive answer ("I consider it likely that Péter moved to Mari's,"; and as for *vajon*, the implicature translates (roughly) as "I wonder...".

As is illustrated in Figures 5 and 6, thus, picking out the sentence type can be regarded as a "basic settlement" of the intensional profile, relative to which discourse particles are responsible for "fine-tuning" it—as well as such further grammatical elements as epistemic modals [15] [45] [47], evidentials [24] [38], miratives [40], special stress patterns, and Austin's [23] expositive verbs as they are construed in Oishi's approach [24]. The following section is devoted to the illustration of these grammatical clues and their pragmatico-semantic contribution to intensional profiles in the ℜeALIS framework.

IV. The ℜeALIS Pragmatics

First of all, let us overview the requirements towards an up-to-date comprehensive pragmatics on the basis of a current paper by Labinaz and Sbisá [38], who argue that we need to go back to Austin's [23] [36] and Searl's [75] [76] speech act theory in order to provide a framework in which all the elements highlighted by the current accounts can be collected and coordinated.

A. Towards a More Comprehensive Speech-Act Theoretical View

According to Austin [23], every speech act comprises three distinct acts: the act of saying something (the locutionary act), what one does in saying it (the illocutionary act) and what one does by saying it (the perlocutionary act). Acts such as assertions, claims, guesses are illocutionary acts and have distinct, albeit related, illocutionary forces. It is worth picking up the idea that illocutionary acts can be described as the performance of a socially accepted procedure, and that this procedure is to lead to a "conventional" effect. Here are some of the elements that a procedure of such a kind should comprise: (i) kind of person that can execute the procedure; (ii) circumstances in which it is appropriate to execute the procedure; (iii) linguistic forms to be used in order to execute the procedure or make it recognizable; (iv) effects of the procedure on the deontic statuses of the participants; (v) appropriate psychological states of the participants; (vi) appropriate subsequent behavior of the participants, to which they are committed by the procedure executed.

B. Another Speech-Act Theoretical View with the Same Aims

Oishi [24] also intends to revisit and develop Austin's speech act theory, to put forward the idea that expositive verbs bring about effects on the on-going discourse, and that evidentials and epistemic modals play discursive functions by indicating those acts. She argues that to indicate (i) the information source of a thing, event, or situation by an evidential, and (ii) the speaker's epistemic attitude toward it by an epistemic modal is to indicate what illocutionary act the utterance performs. Especially, evidentials and epistemic modals indicate a particular type of expositive illocutionary act, which is one of Austin's categories of illocutionary acts. We intend to complete this list of indicators with miratives (e.g., *gee* in English and its Hungarian counterpart *jé*) and with special stress patterns, beyond the choice of the sentence type itself; see Subsection IV-C.

Oishi [24] argues that in performing one of the various types of expositive act, the speaker expounds her/his communicative engagement with the hearer, while inviting him to react to it in a specific way. There are various types of communicative activities that the speaker can provide: in saying an utterance, the speaker does something with a thing/event/situation in the world, with a statement, with the hearer, with knowledge about a thing/event/situation in the world, with the statement that has been imported, and/or with a thought.

All this can be captured in the theory proposed by Oishi [24] in a surprisingly simple way: in the case of each speech act, the speaker is to be distinguished from the addresser of the act, and the hearer from the addressee of the act, and the situation from the context of the act. The dynamism of performing the illocutionary act and the corresponding perlocutionary act, thus, is explained as complex interrelations between the speaker and the addresser, the hearer and the addressee, and the situation and the context.

C. Checking The Complex Pragmatic Interrelations via Generalized Truth-Conditional Evaluation in ℜeALIS

We claim that checking all elements of the aforementioned complex interrelations, as well as the similar ones listed in Subsection IV-A as (i-vi), essentially boil down to the evaluation of truth values over the ℜeALIS universe, defined in Section III, due to the simultaneous presence in a single model of the external world and all of its mind-internal (finite) alternatives; see Fig. 1 in Section II.

As for checking, for instance, whether the particular speaker and the particular hearer are suitable for playing the roles of the addresser and the addressee in the case of certain speech acts, a finite set of external relations must be checked even in the world's most complicated system of speech levels and honorifics in Korean; see Table I.

The plain level, for instance, is used typically by any speaker to any child, to his own younger sibling, child, or grandchild regardless of age, or to one's daughter-in-law, or between intimate adult friends whose friendship started in childhood. The intimate level is used by a child of pre-school age to his or her family members including parents, or between close friends whose friendship began in childhood or adolescence. It may also be used to one's adult or adolescent student, or to one's son-in-law. The familiar level is slightly more formal than the intimate level, typically used by a male adult to an adolescent such as a high school or college student or to one's son-in-law, or between two close adult friends whose friendship began in adolescence. The remaining three levels are used only to adult hearers.

Table I. Honorification in Korean [77] [78]

Clause type/speech level	Declarative	Interrogative	Imperative	Propositive
Plain	*po-n-ta* see-IND-DECL	*po-ni* see-Q	*po-a-la* see-INF-IMPV	*po-ca* see-PROPOS
Intimate	*po-a* see-INF			
Familiar	*po-ney* see-DECL	*po-na* see-Q	*po-key* see-IMPV	*po-sey* see-PROPOS
Blunt	*po-o* see-BLUNT			—
Polite	*po-a-yo* see-INF-POLITE			
Deferential	*po-p-ni-ta* see-ADD.HON-IND-DECL	*po-p-ni-kka* see-ADD.HON-IND-Q	*po-si-p-si-o* see-SUBJ.HON-ADD.HON-REQUEST-IMPV	*po-p-si-ta* see-ADD.HON-REQUEST-PROPOS
Neutral	*po-n-ta* see-IND-DECL	*po-n-unya* see-IND-Q	*po-la* see-IMPV	*po-ca* see-PROPOS

Probably the most popular level is the polite level, which is the informal counterpart of the deferential level. While deferential level is usually used by males, the polite level is used widely by both males and females in daily conversations. Both the polite and the deferential levels are used to a socially equal or superior person, but in general, the polite level is favored between close persons. It is also worth mentioning that no propositive form is available in the superpolite level, presumably due to the fact that the addressee must be so vastly superior when using this level that the speaker could not propose to share an action; and so on.

The details are irrelevant here. What is relevant is that however intricate this system is, capturing the interrelations is nothing more than a typical programming task for ambitious Prolog beginners.

Checking whether a situation is a suitable context for a particular expository speech act, predominantly requires the pattern matching techniques developed in (S)DRT [22], underlying ℜeALIS. Antecedents of pronouns and definite noun phrases are to be sought either in certain worldlets in certain minds (in the course of anaphora resolution) or in the external-world model (in the case of deixis). Setting up rhetorical relations between a new infon and a salient infon in the context also belongs to the task discussed in this paragraph; this is exactly the specialty of SDRT [71] [26].

We conclude this section by noting that what Oishi [24] must consider a crucial difference between expositive acts and the other four classes of illocutionary acts proposed by Austin [23] is no difference in ℜeALIS, in an advantageous way. "They are not changes in the world ... or changes in the social world ... In the expositive act, the speaker and the hearer are assumed to be discourse participants who sustain and develop the discourse. By performing this type of act, the speaker tries to control the discourse [and not the world] and influence the hearer as a discourse participant." In ℜeALIS, the internal contents of human minds are defined as parts of the world—as they are, indeed.

D. Linguistic Clues in Hungarian Sentences and Intensional Profiles

It is high time to exemplify the linguistic elements that contribute to the decision of the speech act, that is, essentially, to that of the intensional profile, in some way or another, as was promised in the last paragraph of Section III. Note that the annotations follow the conventions used in the series *Approaches to Hungarian* (Amsterdam: Benjamins).

The (a) sentence in Fig. 7 is ambiguous due to the modal suffix in italics (*-hAt*). The epistemic (a') interpretation differs from the deontic (a") interpretation in two crucial points [15]. While in the latter case the addresser commits her-/himself to the truth of the infon carrying the piece of information that Péter moved to Mari ('B^M' refers to 'maximal belief', that is, to certain knowledge), in the former case (s)he has only "some certainty". The other difference is that according to the deontic interpretation there is a person r* whose intention it depends on whether Péter can move to Mari. It is asserted in the case in question that r*'s intention towards the move is neutral ('0'), and not positive or negative; r*, thus, permits the move in question, by withholding his intention. In the case of the epistemic interpretation, what the addresser does not intend ('$\langle$I,M,AR,τ,0$\rangle$') is to prevent the addressee from conjecturing ('$\langle$B,sm,ae,τ^{+},+$\rangle$') that Péter moved to Mari.

Sentence (7b) is also a conjecture expressing the addresser's epistemic uncertainty (7b'), but its meaning has an interesting evidential component (7b"): it is implied that the addresser probably saw the action in question.

Sentence (7c) exemplifies a mirative marker, whose peculiar meaning contribution is that the addresser accepts the given infon (7c'), which is new to her or him (7c''), and somewhat surprising (7c''').

a. Péter Marihoz költöz*het*ett.
P. M.Ade move.may.Past.3Sg
'Péter may have moved to Mari's.' 'Péter was allowed to move to Mari's.'
a'. ⟨B,sm,AR,τ,+⟩ a''. ⟨B,M,AR,τ,+⟩
⟨I,M,AR,τ,0⟩⟨B,sm,ae,τ+,+⟩ ⟨I,M,r*,τ,0⟩

b. Péter *mintha* Marihoz költözött *volna*.
P. as_if M.Ade move.Past.3Sg be.Cond.3Sg
'Péter may have moved to Mari's. I may have met them—but I cannot remember.'
b'. ⟨B,sm,AR,τ,+⟩ b''. ⟨SAW,sm,AR,τ,+⟩

c. *Jé*, Péter Marihoz költözött! d. Péter *persze* Marihoz költözött.
Gee P. M.Ade move.Past.3Sg P. of_course M.Ade move.Past.3Sg
'Gee, Péter moved to Mari's.' 'Péter moved to Mari's, of course.'
c'. ⟨B,M,AR,τ,+⟩ d'. ⟨B,M,AR,τ,+⟩
c''. ⟨B,M,AR,τ⁻,0⟩ d''. ⟨B,aM,AR,τ,+⟩⟨B,gr,ae,τ⁻,+⟩
c'''. ⟨D,sm,AR,τ,+–⟩

Figure 7. Markers of epistemic and deontic modality, and evidential and mirative markers in Hungarian

The discourse particle *persze* (cca. 'of course') is used in contexts when speaker and listener share or are supposed to share common knowledge [80]. We claim that it corresponds to the Latin *scilicet*, which "indicates that the evidence is based on expectation ('as is to be expected,' 'of course') and is strongly directed towards the addressee, [in contrast to, say] *videlicet* [which] indicates that the evidence is inferable from the context or reasoning ('clearly') and is not directed towards the addressee" [81]. This special meaning component is captured by the level label presented in (8d'') via referring to an information state of the addressee that precedes the utterance time (τ^-); it is expressed in this way that the addressee could have been almost sure that Péter had moved to Mari without having been informed about that. That is, this almost certain piece of knowledge is not due to any inference drawn on the basis of what has been said by the addresser, but rests upon the given addressee's peculiar knowledge.

Now we provide a comparative overview of different Hungarian sentence types with imperative verb morphology given in (8a-f), whose intensional profiles are presented in the table. Our pragmatico-semantic analyses are chiefly based on Szücs's empirical observations and systematization [82] but other results are also considered [71] [83].

a. Költözzön Péter Marihoz! b. *Kööötöl*tözzön Péter Marihoz!
move.Imp.3Sg P. M.Ade *move*.Imp.3Sg P. M.Ade
'Péter should move to Mari's.' 'Péter can move to Mari's, I do not mind.'

a'. ??Költözzek Marihoz! b'. ??*Kööötöl*tözzek Marihoz!
move.Imp.1Sg M.Ade move.Imp.1Sg. M.Ade
'Let Péter move to Mari's.' 'Let me move to Mari's.'

c. Hadd költözzön Péter Marihoz! c'. Hadd költözzek Marihoz!
let move.Imp.3Sg P. M.Ade let move.Imp.1Sg. M.Ade
'Let Péter move to Mari's.' 'Let me move to Mari's.'

d. [Hadd pletykáljanak]$_e$, [odaköltözöm Marihoz]$_{e'}$!
let gossip.Imp.3Pl there.move.1Sg Mari.Ade
'Let there be gossip, I do not mind, I will move to Mari's.'

e . Költözzön *csak* Péter Marihoz! f. Költözzön *már* Péter Marihoz!
move.Imp.3Sg only P. M.Ade move.Imp.3Sg already P. M.Ade
'Let Péter move to Mari's.' 'I want Péter to decide to move to Mari's at long last.'

Figure 8. The intensional profiles of some Hungarian sentence types with imperative verb morphology (see also Table II)

It is common in all types (see the first two rows of Table II) that the addresser of the chosen speech act is sure that the result phase $\varphi_{res}(e)$ [65] of the given eventuality e does not hold (i.e., Péter and Mari still live in different flats, that is, Péter has not moved to Mari yet) and more or less assumes that the addressee is also aware of this fact (the certainty of her or his assumption is given as 'nM', that is, 'non-maximal').

By performing the basic imperative type (8a), the addresser longs for the aforementioned result state $\varphi_{res}(e)$ and wants the addressee to intend the action e. The addressee's stimulated intention is optimally efficient if (s)he coincides with the agent of the action ("Move to Mari's."). It is, however, definitely excluded that the addresser and the Agent coincide (8a') [84]. Note that (8a) is the only imperative in the strict sense proposed in [71], (8b-f) can rather be classified as proto-imperatives.

Table II. Imperatives and Proto-Imperatives in Hungarian

	a. Basic	b. CVVVC...	c. *hadd*$_1$	d. *hadd*$_2$	e. *csak*	f. *már*
AR's knowledge conc. $\varphi_{res}(e)$	⟨B,M,AR,τ,–⟩	←	←	←	←	←
ae's knowledge conc. $\varphi_{res}(e)$ (acc. to AR)	⟨B,nM,AR,τ,+⟩ ⟨B,M,ae,τ,–⟩	←	←	←	←	←
AR's, ae's and/or Ag's desire conc. $\varphi_{res}(e)$	⟨D,M,AR,τ,+⟩	⟨D,M,AR,τ,0–⟩	⟨D,M,AR,τ,0+⟩	*hadd*$_2$ ←, but for e'	⟨D,nM,AR,τ,–⟩	⟨D,M,AR,τ,0+⟩
		⟨B,nM,AR,τ,+⟩ ⟨D,M,r*,τ,+⟩	⟨B,nM,AR,τ,+⟩ ⟨D,M,ae,τ,0–⟩	For e: ⟨B,nM,AR,τ,+⟩ ⟨D,M,r*,τ,0–⟩	⟨B,nM,AR,τ,+⟩ ⟨D,M,r*,τ,+⟩	⟨B,nM,AR,τ,+⟩ ⟨D,M,r*,τ,+⟩
			⟨B,nM,AR,τ,+⟩ ⟨D,M,r*,τ,+⟩	⟨D,M,AR,τ,0+⟩ ⟨D,s,AR,τ,–⟩		⟨D,M,AR,τ,+⟩ ⟨I,M,r*,τ,+⟩
AR's intention conc. e and/or ae's intention	⟨I,M,AR,τ,+⟩ ⟨I,M,ae,τ+,+⟩	⟨I,sm,AR,τ,0⟩ ⟨I,M,AR,τ,–⟩ ⟨I,M,ae,τ+,–⟩	⟨I,M,AR,τ,–⟩ ⟨I,M,ae,τ+,–⟩	⟨I,M,AR,τ,+⟩	⟨I,M,AR,τ,0–⟩ ⟨I,M,ae,τ+,–⟩	⟨I,M,AR,τ,+⟩ ⟨I,M,ae,τ+,+⟩ ⟨I,M,r*,τ++,+⟩
Note	Ideal: ae = Ag	Preferred: r*= [ae > Ag]	c'. Preferred: AR = r*	Preferred: r*=ae	Preferred: r*= [Ag > ae]	Preferred: r*= [Ag > ae]
	Excluded: AR = Ag	Pref'd/Excluded: r*= ae = Ag ≠ AR	Excluded: ae = Ag			Pref'd/Excluded: ae = Ag ≠ AR

Type (8b) differs from the basic type only in intoning the first syllable of the verb stem in a peculiarly lengthened way. The effect is that now it is not the addresser who longs for the given action but the addressee or the agent of the action. As for intentions, the addresser remains neutral, and does not want the addressee to do anything against e. It is, again, definitely excluded that the addresser and the Agent coincide (8b').

Type (8c) is associated with a third "distribution" of intentions among the three straightforwardly interested participants: the addresser, the addressee and the agent of the action in question. Now it is the addressee who is assumed not to long for $\varphi_{res}(e)$ while the addresser and the Agent long for it. The latter two participants preferably coincide (8c') while this time it is the coincidence of the addressee and the Agent that is excluded.

In the speech act defined by (8e), the addresser is definitely against $\varphi_{res}(e)$, which is now assumed to be longed for very much by the agent (or, perhaps, the addressee). (8f) presents a new distribution of intentions again: the addresser thinks that someone, preferably the agent, longs for $\varphi_{res}(e)$ very much, and (hence) wants this person to realize her or his wishes.

Instead of entering into further details, which obviously does not belong to the aims of this paper, we would like to call the reader's attention to the following generalization: checking whether the speaker, the hearer and the given situation are suitable for serving as the addresser, the addressee and the context of the linguistically defined speech act (à la Oishi [24], see Section IV-B) simply requires a truth-conditional investigation primarily into the addresser's mind's certain worldlets (e.g., what (s)he longs for and assumes certain other persons to long for). The task boils down to get to the worldlets in which certain polarity values must then be checked.

Let us now suppose that a speaker performs (9a), which is the same as (8b), while thinking what is described in (9b). As now all the beliefs, desires and intentions are compatible with the speech act determined by the imperative sentence type and the peculiar intonation, the speaker proves to be impeccably sincere while performing (9a). She or he is undoubtedly suitable for playing the addresser's role in the speech act (s)he has initiated.

a. Speaker: "*Kööölt*özzön Péter Marihoz!"

b. Facts $\prec$ beliefs: $\langle B,M,MY,\tau,-\rangle$; $\langle B,M,YR,\tau,-\rangle$;
$\prec$ desires: $\langle D,sm,MY,\tau,-\rangle$; $\langle D,sm,YR,\tau,+\rangle$; $\langle D,sm,r_{Péter},\tau,+\rangle$;
$\prec$ intentions: $\langle I,sm,MY,\tau,0\rangle$; $\langle I,sm,YR,\tau^{+},+\rangle$

c. Facts $\prec$ beliefs: $\langle B,M,MY,\tau,0\rangle$; $\langle B,M,YR,\tau,+\rangle$;

c'. Facts$\prec$desires: $\langle B,M,MY,\tau,+\rangle\langle D,M,YR,\tau,0\rangle$; $\langle B,M,MY,\tau,+\rangle\langle D,M, r_{Péter},\tau,-\rangle$

Figure 9. Ideal matching between interlocutors and speech act participants, and mistake or deception in matching

Let us now suppose, however, that the relevant facts are as in (9c). That is, the speaker does not know whether Péter lives together with Mari, and the hearer definitely knows that they already have lived together. In a case like this, the speaker was insincere due to her or his bluff, by which (s)he pretends as if (s)he were sure that Péter had not moved to Mari yet. As for the hearer, it would be pertinent to inform the speaker about the (supposed) mistake according to which Péter and Mari still live in different flats. It is in this way that the hearer should get rid of the incompatibility between the speaker's declared presupposition and the real facts in the world.

If the speaker thinks as follows while performing (9a): "you do not bother whether Péter moves to Mari or not, and Péter wants to definitely refrain from moving to Mari," the utterance contains insincerity and/or hypocriticality. The speaker wants to deceit the hearer in respect of her or his knowledge on the hearer's and Péter's wishes.

The intricate phenomena of politeness can also be captured on the basis of the triplets of the readily comparable intensional profiles of the utterance, of the addresser and of the addressee in ℜeALIS (the theory) and by means of ℜeALIS2.1 (the software application). As is illustrated in Fig. 10, the addressee her- or himself must consider two intensional profiles simultaneously in order to be capable of correctly decoding the addresser's real intention. Utterance (10a) is itself ambiguous (10b-c), but it is also worth taking potential ulterior motives into account (10d). The addressee, thus, needs to continuously set up hypotheses during the on-going discourse in the form of intensional profiles, which are worth being compared to the linguistically encoded intensional profiles that belong to the clauses performed by the addresser.

a. Főznél egy levest?
Cond.2Sg a soup.Acc
b. literal meaning: 'Do you want / [feel like] to cook a soup?'
c. request: 'Cook a soup, please.'
d. Suppose the sentence is used in a situation in which the addressee has been telling a story to a group of people, which is unpleasant to the addresser; the addresser intends to interrupt the addressee by this seeming request (or inquiry), in the hope that (s)he forgets to continue the unpleasant story.

Figure 10. Literal meaning, politeness or ulterior motives?

In the particular case, thus, the hearer needs to decide on the basis of her of his own intensional profile about the speaker's beliefs, desires and intentions whether (s)he is really interested in the addressee's wishes (10b), or longs for a good soup (10c), or both hypotheses mentioned can be excluded and/or ulterior motives can me assumed. Observe that checking all these hypotheses can be modeled in ℜeALIS as evaluating polarity values in appropriate worldlets in worldlet-based representations of human minds.

Note in passing that certain polite forms do not yield ambiguity. The form presented in (11c), for instance, used typically by young people to old people, chiefly to old ladies, has no any kind of literal meaning. It is conventionalized in such a way that is to be regarded as an unambiguous expression of a certain speech act.

a. Láttad? b. Látta? c. Tetszett látni?
see.Past.2Sg see.Past.3Sg like Past.3Sg see.Inf
'Have you$_{Sg}$. seen it?'

Figure 11. Degrees of politeness in Hungarian

The type illustrated in (11b), however, is ambiguous between a "literal meaning" with a 3Sg. subject ('Have (s)he seen it?') and a polite interpretation in which the subject is the addressee.

V. USERS AND USES

ℜeALIS1.1 (just like ℜeALIS2.1) is worth demonstrating as a software application intended to supply collaborating linguists and certain types of non-linguist experts, because in this way it is quite easy to capture, at the same time, the required input data types and the output production.

A. Internal Users

Our software application ℜeALIS1.1 is primarily intended to supply linguists with a device to build fragments of arbitrary languages of arbitrary morphological types (NB: it is worth elaborating applications based on Hungarian since this language is an ideal challenge due to its extremely rich morphology [18] [20] [55]). These fragments can capture such specialties of human languages as, for instance, the compositional cumulation of meaning units [25]. The definable meanings are pragmatico-semantic descriptions that satisfy the relevant definitions of ℜeALIS. The group of users defined in Section V-A will be referred to as internal users.

B. External Users

Those using the developed language fragment will be referred to as external users. In the course of using the software application, they can select lexical items to build sentences, the (generalized) truth-conditional interpretation of which they will be given on the basis of a "multiplied world model", which they have themselves constructed or received from "internal" experts [12].

Possible external users may be detectives or judges, for instance, who can have the truth of groups of propositions evaluated. In harmony with our "constructionist" stance, we mean by the aforementioned "generalized truth-conditional evaluation", besides the final *true/false* value, the collection of all the information required to reach this truth value. Our software application thus, among others, serves the purpose of collecting and systematizing data in the effective structure that ℜeALIS offers. ℜeALIS2.1 offers an even more extended truth-conditional evaluation, which pertains to all the pragmatic aspects discussed in Section IV. It is checked, thus, whether the speaker, the hearer and the given situation are suitable for serving as the addresser, the addressee and the context of the linguistically defined speech act, on the basis of the model of the external world and its human-mind-internal images called worldlets in ℜeALIS (Section IV-D).

C. Defining Relations

Internal users can define an external world w_0, over the universe of which (consisting of entities u_i) they can define relations of different arities [25]. One argument of all these relations is to be a series of disjoint temporal intervals. The program is to "dictate" (through permanent queries) the development of the external world: it requests new and new relations, and in the case of a given relation it requests the provision of (the initial and final points of) temporal intervals (among others).

Such relations can be defined in this way that are homogeneous in the sense that they qualify as true or false "momentarily", i.e., at each internal point of the temporal intervals independently. In Hungarian, for instance, *utazik* 'travel' and *úszik* 'swim' are homogeneous relations while *hazautazik* 'travel home' and *átússza* 'swim across' are heterogeneous. Further, each argument position of a relation can be associated with other relations of the group of relations defined earlier that provide us with restricting information. The agent argument of the Hungarian verb *utazik* 'travel', for instance, can be associated with the restricting relation *ember* 'human'.

D. Defining Label Strings of Worldlets

Relative to the set of "worldlets" (small partial models of alternative worlds) defined up to a certain point, the internal user can define (by simultaneous recursion) a new worldlet where the basis of this definition is the singleton consisting of the external world w_0. Specifically, relative to a worldlet w', a worldlet w" can be determined through a quintuple of labels like the one shown in (12a). Recall that it defines the worldlet containing a human being's (r_{Sue}) knowledge ("maximal" belief); see sentence (2b) in Section II and (4a) in Section III.

a. $\langle \text{BEL,MAX,r}_{\text{Sue}},\tau'',+\rangle$
b. $\varnothing$ / + / – / 0 / 0– / 0+ /+– / 0+–
c. $\langle\langle \text{BEL,sm,r}_{\text{Joe}},\tau,+\rangle\langle \text{INT,MAX,r}_{\text{Sue}},\tau',+\rangle\langle \text{BEL,MAX,r}_{\text{Joe}},\tau'',+\rangle\rangle$

Figure 12. Labeling worldlets.

Alternatives to label BEL are labels INT (intention) and DES (desire), among others. Alternatives to label MAX are lower levels of intensity: e.g., aM (almost maximal). The fourth member of the label quintuple is polarity; the values of this parameter are listed in (12b), on the basis of formula (1) in Section III.

The software application can show through what kind of defining steps one can reach a worldlet relative to the external world as a fixed starting point. The label string in (12c), for instance, defines a worldlet that is to be regarded as the collection of information the status of which can be captured, for instance, by means of the linguistic expression shown in (2c) in Section II: "Joe guesses that Sue definitely wants to convince him to take it for granted that [...]." The worldlet where we should get, thus, is inside Joe's mind, immediately embedded in a worldlet containing thoughts that Joe attributes to Sue. The label of the worldlet in question expresses that it consists of Sue's assumed intentions towards exactly Joe himself.

E. Worldlets, Infons and Polarity Values

Internal users can assign pieces of information to worldlets. This procedure is to be "dictated" by the program as follows.

In the more general case, a point in time must be specified. As a reaction of the program, on the basis of the above-discussed temporal-interval series belonging to the relations, it is written which relations stand between which entities at the given point of time. If the user specifies,

besides a point in time, a relation and some entities that occupy certain argument positions of the relation, the task of the program remains the writing of the lacking entities that stand in the given relation with the provided entities at the provided point in time. The unit of this writing process is the external infon [61]: an infon means the piece of information that certain entities stand in a certain relation at the given moment (e.g., Joe loves Sue, or Joe is just traveling).

Internal users can assign an infon (produced in the way sketched above) to an arbitrary worldlet for an arbitrary temporal interval. The application of this temporal interval serves the purpose of capturing such factors as the dwindling into oblivion or some re-categorization of pieces of information.

Assigning a group E of infons to a worldlet standing with the external world in the relation provided in (13a) can be interpreted as follows: Sue perceives information E from the external world and accepts as the current state of her environment. A similar interpretation in the case of the complex relation provided in (12c) is as follows: Joe suspects that Sue wants to make him to be sure that information E is true (while Sue herself, for instance, does not necessarily believe in the truth of E; nor is E true in the external world).

If the same infon is simultaneously assigned to someone's positive belief-worldlet (see '+' in (12b)), negative desire-worldlet ('–') and neutral ('0') intention-worldlet, this complex "evaluation" captures this typical situation: the person in question perceives something and accepts its truth, but longs for its opposite without intending to change it (at least at that moment).

It is worth noting in connection with the polarity values listed in (12b) that if an entity does not stand in the relation 'be a linguist' in the external world, then the infon declaring the given entity's momentary being a linguist is to be assigned to the experiencer's negative ('–') or 'undefined' ('∅') belief-worldlets depending on the restricting relations, mentioned in Section VI-A. Ben, for instance, can be thought by an experiencer to be "not a linguist" while in the case of the Eiffel Tower, its being a linguist is undefined.

As for the combination +– in (12b), see (4b) in Section III, together with the relevant comments there. Further uses of the polarity-value combinations are exemplified in (5d), (6c-e), (7c'''), (8b-f) in the previous sections.

F. Information Not Coming from Outside

Internal users can also assign information to worldlets indirectly, that is, not on the basis of (the relations of) the external world. This is "dictated" by the program as follows.

The program asks for predicate names and argument numbers, and then produce argument places with inserted "new" entities, which the software must also urge the user to anchor to "old" (external or internal) entities (NB: their anchoring to any entities is only a possibility). Section V-E, where we defined the procedure of creating infons assigned to worldlets in human minds on the basis of states of affairs in the external world, is worth completing with a short comment. An internal infon does not contain the same entity names as the corresponding external infon does. Instead of identifying them, the correspondence between external and internal entities is accounted for by (α-) anchoring elements of the former group to those of the latter group. In this way, we can explain the cases of misunderstanding where the same external fact is linked to different participants in two experiencers' minds (see the definition of ℜeALIS in Section III).

G. Building the Lexicon

The internal user is given a core lexicon on the basis of the predicates the creation of which was described in Section VI-C; and this core lexicon is enriched with the predicates created in the way described in Section VI-D. Elements of the latter group of predicates must be associated with meaning postulates [25], by the help of queries of the program.

Note that items of the core lexicon need not be associated with meaning postulates since their interpretation is trivial on the basis of their creation: as they have been created by copying certain "patterns" of the external world, the rule concerning the pattern matching their semantic evaluation is based on is automatic. True perception and pattern matching is the same process, considered from opposite directions (cf. Searl's world-to-word and word-to-world directions of fit [75]).

Let us return to the predicates whose forms are defined in Section VI-D; they are assigned meaning in the way to be defined in Section VI-F. Before entering into details, it must be noted that this is the crucial innovation of ℜeALIS1.1, because this is the toolbox which exploits the advantages and results of all the model-theoretic theories, the discourse-representational innovations and the proof-theoretic ideas, and the "diagnosis" of cognitive linguistics on the weaknesses and shortcomings of these three approaches.

What comes from formal semantics [25]? The procedure of pattern matching. Further, the application of interpretational bases used as alternatives to each other ("possible worlds" $\rightarrow$ ℜeALIS-worldlets). And the consideration of the rate of successful instances of pattern matching compared to the entire set of possible instances of pattern matching.

The idea of operation over the partially ordered system of worldlets is due to DRT [21] [22]. The step-by-step execution of this operation, referred to as 'accommodation' in DRT, coincides with the proof-theoretic processing of semantic information [43].

The modeling of the following linguistic elements is basically due to cognitive linguists [48] [85]: *me*, *you*, *(s)he*, *here*, *there*, *now*, *then*, *these here* (in the context), *those there* (demonstration); see also [38].

H. How to Define Lexical Items

The program must help the internal user (the formal linguist) in assigning (groups of alternative) phonetic forms and meaning postulates to predicate names, besides such straightforward information as (sub)categorization and argument number.

Meaning postulates essentially consist of first-order formulas. The most peculiar element of our method is that

each formula like this must be associated with a set of such chains of worldlet labels as the one shown in (12b) and the information as to which worldlet(s) these chains to be linked to in the course of interpreting sentences (possibilities are the external worldlet, certain worldlets of the selected speaker/addresser, hearer/addressee, or participants referred to in the sentences, or worldlets that can be identified in the selected context or scope of demonstration (see the last paragraph in Section V-G).

I. Use Cases for External Users Building the Lexicon

The external users—who can construct a sentence and specify the speaker, the hearer, the entities assumed to be present in the context (and possibly a subset of those in the scope of some demonstration), the speech time and the time of reference, among others—are given a generalized truth evaluation. This means that they are given not only a truth value but also all the pragmatic well-formedness conditions of the sentence "performed" in the specified situation to be construed as the context of a certain speech act.

Thus, they can look "inside" all relevant participants' minds (i.e., the current and possibly some previous information states). They can realize, for instance, whether the definite noun phrases are suitable for unambiguously identifying the intended *denotata*. They can also receive information about the success of satisfying other kinds of presupposition. They can detect, through comparing the information provided by the sentence and the information found in the specified interlocutors' appropriate worldlets, whether there might emerge some misunderstanding, lie, bluff, deception [5], as is expounded in Section IV-D.

VI. LINGUISTIC EXAMPLES

External users are given a peculiarly multiplied data base that contains, besides a relational model of some fragment of (the history of) the external world, several of its alternatives. Recall that these alternatives essentially play the role of possible worlds, known from intensional model-theoretic semantics, but they are finite constructions appearing as such parts of information-state models of interlocutors that can be construed as their beliefs, desires, intentions, or any other kinds of fictions.

A. Generalized Truth Evaluation

The above-sketched arrangement of worldlets enables us to carry out truth evaluation not only on the basis of the external world, which is necessary and sufficient, for instance, in the case of sentences (13a-a'), but also on the basis of internal worldlets, which is obviously necessary in the case of sentences like (13b). The truth of the variants shown in (13b) does not depend on any facts in the external world. It depends on nothing else but Joe's knowledge, or the knowledge that Sue attributes to Joe. In this latter case, it requires more steps to reach the worldlet that can serve as the basis of truth evaluation (external world model → Sue's belief → Sue's hypotheses on Joe's beliefs); cases like this make it necessary to localize worldlets in the recursive way illustrated in (12c) in Section V-D. Verbs expressing modal attitude (e.g., *think*, *guess*, *conjecture*, *wish*) and many other expressions (e.g., *according to someone*) can be associated with meaning postulates by means of the tool described in Section V-G: the essence of their meanings lies with the "direction indicator" function. Such direction indicators help us finding the worldlets that can serve as the basis of the truth evaluation of the proposition that appears in the appropriate argument positions of the modal verbs or other linguistic expressions in question.

a. It was snowing. a'. It has snowed.
b. (Sue thinks that) Joe knows that it was snowing.
c. Patty was traveling home.
d. *That tall Finnish woman* is pretty.

Figure 13. Generalized truth evaluation relying on worldlets.

B. Past Continuous and Present Perfect

Internal users can work out exacting and sophisticated syntaxes and semantics by the help of the toolbox offered by ℜeALIS1.1.

The truth value of (13a), for instance, can be calculated in the following way: the program must query the values of *then* and *there*, and then it localizes the area of the temporal external world model where pattern matching is to be attempted in order to decide whether it is snowing "then" and "there".

The truth evaluation of (13a'), however, requires the values of *here* and *now*, and what is to be checked in the external world is whether the landscape is snowy. The meaning postulate of the verb *snow*, thus, contains the determination of the result state (*snowy*), too, discussed in detail in [65], for instance. Note in passing that (13a') pragmatically suggests that it is not snowing at the relevant moment while the land is snowy as a result of an earlier snowing.

C. Progressive Aspect

The truth evaluation of sentence (13c) also requires a polished and exacting meaning postulation because not only facts of the external world is to be taken into account. A progressive sentence like this is also to be evaluated to be true in a case in which Patty never got home (this observation is called the Imperfective Paradox [86]) but she proves to have been travelling at the moment of *then*, she proves to intend to come home, and the speaker proves to attribute a quite high likelihood to this arrival (Section V-H) [21] [65]. Thus, the content of certain internal worldlets is to be checked, besides the partial satisfaction of a travelling event in the external-world model.

D. The Intensional Character of Nicknames

A demanded pragmatico-semantic analysis of nicknames also requires the toolbox sketched in Section V-H. Who is *Patty* in (13c), for instance? Internal users can capture the essence of the task of finding *denotata* by construing nicknames as special predicates whose "truth evaluation" involves not (only) the external control on the correspondence between official names and nicknames but (also) the worldlets concerned in the following questions: is *Patty* a possible nickname of the speaker for the given person, does the speaker think that the addressee may (also) call her Patty, do they know this about each other, and so on. Hence, internal worldlets are to be checked via pattern matching.

E. (Partially) Subjective Predicates

Example (13d) illustrates further advantages in meaning postulation of the toolbox demonstrated in Section V-H. The adjective *pretty*, for instance, is worth regarding as a fully personal and subjective judgment, with no extension in the external world. Nevertheless, (13d) does not mean exactly the same as the sentence *I consider her pretty*. The truth of this latter sentence exclusively depends on the speaker while it would be elegant to base the evaluation of sentence (13d) on a somewhat less speaker-dependent calculation. As follows, for instance: (13d) is considered true if *most* persons in the external-world model consider the given lady to be pretty. According to an even more elegant solution, instead of the entire set of persons, only those *respected* by the speaker are considered. The extension of the verb *respect* is to be checked in the external-world model, in a way that is essentially the same as checking the appropriate interpersonal relationships in the case of the superpolite Korean interlocutors discussed in connection with Table I in Section IV-C

F. Demonstration and Anchoring

The demonstrative noun phrase in (13d) illustrates another instance of the necessity for "pragmatically conscious" truth evaluation. *That* asks for the value of the "those there" parameter from the external user. It is elegant to assume that this value is a set of entities, out of which the program selects a unique entity on the basis of the predicates *tall*, *Finnish* and *woman*. Their extensions count in the external world, at least primarily; it is an elegant facility, however, to inspect the speaker's beliefs as well, or the speaker's hypothesis about the addressee's beliefs: sentence (13d) can be evaluated as *true but ill-formed* if, for instance, the speaker intends to refer to a tall Swedish woman about whom they think, incorrectly, that she is Finnish.

VII. ℜEALIS1.1, ℜEALIS2.1 AND RUDI

Three software applications are compared with each other in this section.

A. ℜeALIS1.1

The implementation ℜeALIS1.1 is a client-server Windows application that has been elaborated in a Delphi environment, which guarantees rapid and flexible development. Access to data is executed via standard SQL commands by means of a relational data-base management system. For this purpose, we currently use Firebird Interbase.

The Prolog basis, applied in the experimental phase of our research [58] [59] [11], has been replaced with Delphi environment, which is more capable of managing large data-bases, developing user-friendly interfaces, and constructing more complex applications. This is the radical difference between ℜeALIS1.1 and the aforementioned works (NB: the Prolog basis is retained in certain software applications of ours [54] [87]).

The menu items correspond to the services sketched in Section V. Particular menu items are available to the different kinds of users (also defined in Section V) after checking their identity and authenticity. In the course of parsing sentences, morphological input is produced by the procedures demonstrated in [20] [51] [55] , agreement relations are checked by a method similar to the one shown in [88], dependency relations are calculated by means of our special rank parameters [56], and Prim's algorithm is used for producing one or more spanning trees with a minimal cost. At some points, the program provides illustrations of the structures constructed by either the system or its users: for instance, parsing trees, systems of worldlets, and anchoring relations of entities are illustrated. Fig. 14 presents this last facility.

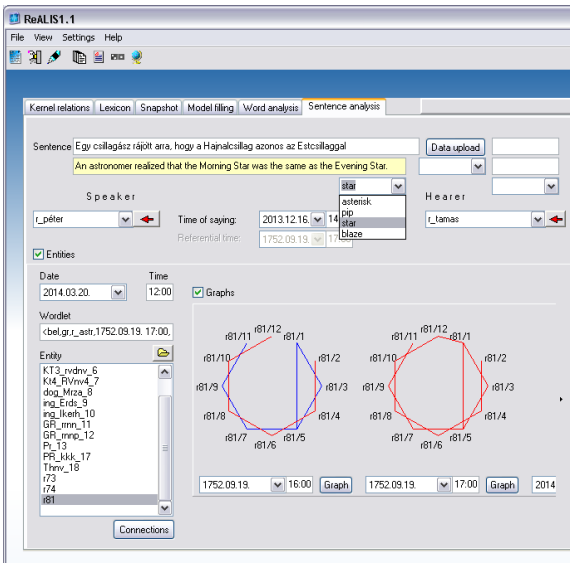


Figure 14. Interpreting the sentence shown in Fig. 3.

The software application is permanently developed and expanded, exploiting new scientific results; it has repercussions on the theory, due to the fact that the theory can be tested by means of the program. We are working on developing tests to evaluate its efficiency.

B. RUDI

As our software application inherently belongs to a radically new and holistic "pragmalinguistics" theory (Section II), it is uneasy to compare to software applications based on some different theoretical foundation. An exception is the SDRT-based experimental software dialogue system, RUDI [26], primarily due to its distinguished attention to the relationship between pragmatic phenomena and the external-world model.

RUDI ("Resolving Underspecification with Discourse Information") computes automatically some aspects of the content of scheduling dialogues, particularly the intended denotation of the temporal expressions, the speech acts performed and the underlying goals. RUDI has a number of nice features: it is a principled approximation of a logically precise and linguistically motivated framework for representing semantics and implicatures; it has a particularly simple architecture; and it records how reasoning with a

combination of goals, semantics and speech acts serves to resolve underspecification that is generated by the grammar.

RUDI analyzes such definite descriptions as requiring a bridging relation to an antecedent in the context. Fig. 15 provides an example: it should be calculated what the temporal expression '4pm' refers to in (the model of) the external world.

A: "Can we meet on Friday?"
B: "How about 4pm?"

Figure 15. Resolving the referential underspecification in the dialogue, which requires revealing the speech act type.

Neither the bridging relation nor the antecedent are determined by the compositional semantics of the utterance, however. Thus, RUDI takes the semantic representation of such expressions to contain an underspecified relation between an underspecified antecedent and the referent for the expression. A task that is co-dependent on resolving this underspecification is computing how the utterance contributes to a *coherent* dialogue. Following SDRT [22], it is assumed that a dialogue is coherent just in case every proposition (and question and request) is *rhetorically connected* to another proposition (or question or request) in the dialogue, and all anaphoric expressions can be resolved. The rhetorical relations are viewed as *speech act types* in the RUDI project—that is the point where ℜeALIS2.1 can be regarded as an extension of RUDI, given that in the ℜeALIS theory further relations among pieces of information stored in (different worldlets of) minds, addressers, addressees, contexts and the external world are (intended to be) taken into account in a completely uniform system (see Section IV-C).

It is also worth noting that the creators of RUDI [26] represent the same holistic stance typical of ℜeALIS implementations in regarding information as flowing either from resolving the semantic underspecification to computing the rhetorical relation, or *vice versa.* They, thus, consider rhetorical relations (that is, practically speech acts) to be an essential source of information for resolving semantic underspecification that is generated by the grammar.

ℜeALIS essentially follows SDRT, which represents discourse content as a "segmented discourse representation structure", which is a recursive structure of labelled DRSs [27], with rhetorical relations between the labels. In contrast to traditional dynamic semantics (see also [27], for instance), SDRT attempts to represent the *pragmatically preferred* interpretation of a discourse—just like ℜeALIS.

The rule schema used in RUDI contrasts with the plan-recognition approach to computing speech acts [89], which uses *only* the goals of the antecedent utterance, rather than its compositional and lexical semantics directly, to constrain the recognition of the current speech act. The dialogue presented in Fig. 16 illustrates the point: it depends on the particular external denotata of the temporal expressions whether interlocutor B has rejected A's proposal or has declared that (s)he is prepared for elaborating it as a common aim.

A: "Can we meet next week??"
B: "I'm busy from the 16th to the 25th"

Figure 16. Computing the speech act (Rejection or Elaboration) requires considering the temporal reference in the external world.

There are a number of advantages, thus, to allow direct access to the content of inferences. The successful performance of the current speech act is often dependent on the *logical structure* of the antecedent utterances, and goals do not reflect this logical structure; rather compositional semantics does (following DRT). An utterance in the context is chosen to which the current utterance can be attached via a rhetorical relation, and this in turn determines which antecedents are available.

C. ℜeALIS2.1 as an Extension of ℜeALIS1.1

It is very important to us that our software applications, which are permanently being developed, have repercussions on the theory, due to the fact that the theory can be tested and sophisticated by means of the program. ℜeALIS2.1, thus, has been being developed (as an extension of ℜeALIS1.1) exactly due to the enormous pragmatic extension of the ℜeALIS theory itself, demonstrated in Section IV.

Nevertheless, this part of the difference between ℜeALIS2.1 and ℜeALIS1.1 is predominantly quantitative, and not qualitative. It had also already belonged to the decisive properties of ℜeALIS1.1 that, based on the truth-evaluating pattern-matching mechanisms between linguistic representations and world models, the same kind of pattern matching is executed everywhere in the uniform system consisting of a model of the external world and a partially ordered conglomerate of its (meticulously labeled) finite human-created images we have dubbed *worldlets*. ℜeALIS2.1 is more developed than ℜeALIS1.1 in the sense that much more relations are evaluated in this intricate database, whose crucial specialty is that one and the same piece of information appears simultaneously in innumerable places in the system (just like images on different sides of a prism).

Let us overview the main problems we should cope with.

Certain difficulties have to do with the fact that we are working on a completely general toolbox that utilizes the afore-mentioned multiplied world model, that is, one that is underspecified in several respects but can be rapidly specified when it is designated for particular purposes. This also holds for the pragmalinguistics input; some difficult grammatical phenomena that must be captured in a demanding way are collected in Section IV, VI and VII-E. Due to the uniform and holistic approach of ℜeALIS, we cannot afford to use parsers or other devices developed in other projects. Elaborating sufficiently sophisticated testbeds, however, require very much cost, time, energy and creativity.

Recursivity is another stubborn problem. Unlimited chains of linguistic expressions can be produced, whose elaborated pragmatico-semantic analysis leads to proliferation problems.

It is also difficult to register the copies of multiplied entities in almost identical alternative models. We need to apply safe and effective but very rapid methods in copying

huge databases in a way that makes it possible for us to carry out the relevant differences between them.

Nevertheless, the most difficult task is the safe and systematic treatment of temporal entities, which come from the model of the external world as well as from the alternative models, and also come from the event structure of lexical items [65] and from the discourse structure of sentences to be parsed. We have been led to the conclusion that the utter key to different kinds of systematization problems is utilizing points of time as "identifying stamps".

Our ultimate task in this area, thus, is no less than modeling the episodic memory of the human mind [38]—together with the semantic memory.

D. ℜeALIS2.1 as a Model of Long-Term Memory

Leiss [50] distinguishes the aforementioned two parts of long-term memory on the basis of Tulving's research [90] as follows: "the semantic memory stores knowledge, whereas the episodic memory stores experiences".

It is relevant that, in contrast to semantic knowledge, episodic knowledge has space-time-index quality, because experience characteristically takes place in space and time, and consequently the respective space-time coordinates are mapped. As a rule, it is assumed that concepts emerge, in that experiential data are generalized so that they no longer contain space-time coordinates. The construction of episodic memory typically correlates with the acquisition of finite sentences, which, in turn, correlates with the use of inflected verbs. The function of finite sentences is to establish a reference by anchoring concepts to a space-time context. Without this technical device, autobiographical memories would not be possible. Space-time coordinates are constituted by the ATMM-complex, mentioned in Section II: by the aspect-coded space coordinates, by the time-coded tense coordinates, by mood coordinates that signal *irrealis* (i.e., statement not anchored in reality) versus *realis* in the sense of Carnap, and by the coordinates of the source of evidence the speaker relies on (see Fig. 7 in Section IV-D). The download of episodic experience by virtue of the grammatical categories aspect, tense, mood, and modality enables us to orient ourselves in the real world. These categories generate a system of coordinates that anchors our activities in the world, and which, in turn, provides indices for our memories, thereby makes them memorizable.

The very difference between semantic memory and episodic memory consists of the fact that experiences are based on the first person, whereas knowledge is based on the intersubjectification of first-person experiences. Inter-subjectification implies the neutralization of space-time coordinates, thus generating knowledge. Subjective certainty, and the download of this type of certainty, is achieved by virtue of the grammatical ATMM-categories. Here the functions of language are essential for gaining reference to the world. Objective certainty will be achieved by the never-ending construction of an intersubjectively negotiated lexicon.

We claim that, due to its lifelong character, the part of the worldlet structure of ℜeALIS where the internal entities (i.e., referents) are anchored to external entities in the world model (i.e., to real objects and persons) can readily be regarded as an implementation of episodic memory. Semantic memory consists of worldlets that contain referents that are not "out-anchored".

E. Missing Links

Due to the aforementioned ℜeALIS-model of long-term memory, the mechanisms captured in Pustejovsky's Generative Lexicon [91] can be implemented. That is, it is possible to derive "a potentially infinite number of senses for words from finite [lexical] resources," and to explain "the interpretation of words in context." This can be regarded as a qualitative innovation in ℜeALIS2.1, compared to ℜeALIS1.1.

The word *London* in (18a), for instance, is used in the given context as an attributive of the noun *train*. The problem with it is that while *a London flat* is a flat that 'can be found in London,' the London train in question is claimed to be in Bristol at the time of reference. In another context, *the London train* may refer to a train that can be found in Manchaster, in a train museum, and the attribute *London* refers to the city in which it was produced, or in which the museum can be found from which the given train has been borrowed. How can then the contextually adequate meaning be calculated?

a. "The *London* train arrived in Bristol."
b. σ relation: ⟨e1,r_{arrive},t1,r11,r12⟩; ⟨e2,r_{train},t2,r21⟩; α: ⟨r21,r11⟩
c. ⟨e3,$r_{in\text{-}London}$,t3,r31⟩; α: ⟨r31,r21⟩
d. σ relation: ⟨e4,r_{go},t4,r41,r42,r43⟩; ⟨e5,r_{train},t5,r51⟩; α: ⟨r51,r42⟩
e. ⟨e',r'_{train},t',r'1,r'2,r'3⟩
e'. ⟨e3,$r_{London.Adj}$,t3,r31⟩; ⟨e6,r'_{train},t6,r61,r62,r63⟩; α: ⟨r31,r62⟩

Figure 17. Adjectives with an unbounded number of meanings.

As was defined in Section III, the σ eventuality function is responsible for "formulating" the elementary statements the given sentence provides, from word to word. (17b) presents that it is claimed that something arrived somewhere, which is a train. What is expressed in (17c), however, is incorrect, assumed that the attributive means 'can be found in London' (as a reasonable primary meaning).

It is at this point that it is worth having recourse to the episodic memory in order to use it as a huge database. Suppose it contains episodic information expressing the fact that once 'a train went from a certain place to a certain place' (17d). If this information is deprived of the external anchors, such a temporary predicate can be constructed for the semantic memory by unifying its parts that has three arguments: 'r'1 is a train from r'2 to r'3' (17e). A copy of this temporary predicate can then be attempted to be applied in the episodic memory again, to whose *from*-argument the-entity-that-can-be-found-in-London (referent r31 in (17b)) can successfully be anchored, as is formulated in (17e'). Thus, the solution is that it is the station from which the given train departed that can be found in London.

Note that it is not excluded that other "solutions" may also come out as results; but it is sure that all results will come out that a human being is capable of finding on the basis of her or his past experiences. The emerging

"competing" results then can be compared with each other on the basis of multiplicity of computing and fitting into the broader context.

A Hungarian ontology has also been built in ℜeALIS2.1 because the above-sketched search for the context-dependent adequate interpretation of the expression *London train* can be made more efficient by attempting to replace *train* with such sister categories as *bus* and *airplane*, for instance, and *London* with *Glasgow* or *Manchester*, since the expression *Glasgow bus* can also help finding the missing link "scheduled service". We are investigating what kinds of substitution can increase efficacy, and what kinds or substitution prove to be definitely harmful.

Fig. 18 presents a verb with an underspecified meaning. In (18a), it seems to take two noun phrases as arguments; and suppose that in the case of an earlier occurrence, the associated specified meaning was 'finished reading (a book).' However, this reading is incompatible with (18a).

a. "Ed *finished* the sandwich."
b. σ relation: ⟨e2,$r_{sandwich}$,t2,r21⟩
c. ⟨e1,$r_{finish\text{-}reading}$,t1,r11,r12⟩; α: ⟨r21,r12⟩
d. σ relation: ⟨e',r'$_{finish}$,t',r'1,e'1⟩
d'. σ relation: ⟨e3,r_{eat},t3,r31,r32⟩; ⟨e4,$r_{sandwich}$,t4,r41⟩; α: ⟨r41,r32⟩
e. σ: ⟨e5,r'$_{finish}$,t5,r51,e51⟩; ⟨e6,r_{eat},t6,r61,r62⟩
e'. α: ⟨e6,e51⟩; ⟨r62,r21⟩

Figure 18. Verbs with an unbounded number of meanings.

It is reasonable to assume that the episodic memory contains, on the one hand, an occurrence of *finish* with an argument position for expressing actions (18d), and, on the other hand, information about a sandwich that has been eaten (18d'). Hence, copies of these pieces of information can be created for the semantic memory, on the basis of which, then, the story can be put together in the episodic memory, again, according to which the "sandwich finished" has been *eaten* (18e-e').

Nevertheless, other solutions may also come out, which may prove to be better in certain contexts. Ed, for instance, might have managed to *butter* the given sandwich. It is also worth mentioning that the aforementioned involvement of an ontology may also have a positive influence in finding the "missing link(s)" in the case of verbs with potentially infinite meanings. The expression sandwich, for instance, is worth attempting to be replaced with *soup* or *cake*, whose eating can also be finished.

The verb *resemble* in (19a) may also be problematic if, say, in the course of its earlier occurrence the associated specific meaning was 'resemble in being remarkably tall.' This meaning is excluded if, say, Ed is tiny.

a. "Ed *resembles* Ted."
b. σ: ⟨e1,$r_{resemble\text{-}in\text{-}being\text{-}tall}$,t1,r11,r12⟩; ⟨e2,$r_{Ed}$,t2,r21⟩; ⟨e3,$r_{Ted}$,t3,r31⟩
c. α: ⟨r21,r11⟩; ⟨r31,r12⟩
d. σ relation: ⟨e',r'$_{resemble}$,t',r'1,r'2,e'1⟩
e. σ relation: ⟨e4,r'$_{resemble}$,t4,r41,r42,e43⟩; ⟨e5,p5,t5,r51,...⟩; ⟨e6,p6,t6,r61,...⟩
e'. α: ⟨e5,e43⟩; ⟨e6,e43⟩; ⟨r51,r_{Ed}⟩; ⟨r61,r_{Ted}⟩

Figure 19. Meta-level expressions?

Such pieces of information need to be found in the episodic memory as those presented in (19d-e); first of all, an occurrence of *resemble* with an explicit *in*-argument for properties. Then predicates must be found there which have held for both Ed and Ted. We hypothesize that what should be found is not a great set of "shared" predicates, but rather a single one that is salient in some way in the context. Note that searching for predicates is not a second-order procedure in ℜeALIS because, due to reification [8], predicates are assigned to Landmanian pegs [62] in the worldlets in the same way as arguments are, so they take part in the σ relations as equal internal entities (cf. [92]).

The temporal adjective presented in (20a) patterns with the last three special words in being used not immediately as a simple predicative element. Thus, it cannot be claimed what is given in (20c), namely, that "he is former," while it can be claimed that "he is a spy," or "he is old." That is why the temporal adjective in question is called an irregular adjective by Kiefer [92].

a. "I met a *former* spy."
b. σ relation: ⟨e1,r_{meet},t1,r11,r12⟩; ⟨e2,r_{spy},t2,r21⟩; α: ⟨r21,r12⟩
c. ⟨e3,r_{former},t3,r31⟩; α: ⟨r31,r21⟩
c'. t2<t1

Figure 20. Irregular adjectives I.: temporal adjectives.

The solution of the "equation system" is quite simple in ℜeALIS, given that ordinary verbs, nouns and adjectives provide σ-formulas containing temporal referents as well. In (20b), t1 and t2 are the temporal referents. If (20a) did not contain *former*, both t1 and t2 would be identical to the reference time. The contribution of *former*, thus, is the ordering between t1 and t2 presented in (20c'). That is, the person in question is claimed to serve as a spy earlier than the time of the meeting.

We conclude this section with another adjective called irregular by Kiefer [93] essentially for the same reason: it cannot be claimed that "someone is *alleged*." The contribution of this adjective to the content presented in (21b), thus, is not (21c).

a. "I met an *alleged* spy."
b. σ relation: ⟨e1,r_{met},t1,r11,r12⟩; ⟨e2,r_{spy},t2,r21⟩; α: ⟨r21,r12⟩
c. ⟨e3,$r_{alleged}$,t3,r31⟩; α: ⟨r31,r21⟩
c'. level λ of e2 (rel. to e1): ⟨B,M,r*,τ,+⟩

Figure 21. Irregular adjectives II.: modal adjectives

The solution of this puzzle has to do with the crucial weapon of ℜeALIS: the contribution of *alleged* is that infon e2 (according to which someone is a spy) must be set in a worldlet different from the worldlet in which e1 is to be set (according to which the speaker met someone). While it is claimed by the speaker, thus, that (s)he met someone, the claim that he is a spy is attributed to another person by the speaker (21c').

VIII. SUMMARY

We have intended to convince the reader that it is not a huge cost for the (theoretical, and then computational) treatment of such "internal questions of language" as sentence types (Sections III-IV), honorification, epistemic and deontic modality, evidential and mirative markers, expression of politeness, special intonations (Section IV), tense, aspect, subjectivity, deixis (Section VI), irregular adjectives, and the problem of deriving a potentially infinite number of senses for words from finite lexical resources (Section VII) to attempt to formally describe information states of human minds in communication—that is, the human mind itself (Section I).

Our ambitious stance can be legitimized by the Un-Cartesian philosophical hypothesis on language, defended by Leiss [38] as follows: "...language is a type of translation of the world into mental representations. Language is a technical device for diagrammaticizing the world. It is not a device that directly reflects the world, but rather a motivated reduction of the complexity of reality through transduction into more or less specified diagrams of the world. Language enables us to
do with the assistance of the technique of grammar (referential opposition of concepts), which enables us to orient ourselves in the real world in space and time. Beyond that, language provides the option, again on the basis of the construction of concepts, to generalize experiential certainties, thereby making them usable in contexts hitherto unexperienced."

This paper demonstrates what kind of ontological innovation (Section II) and mathematical techniques (Section III) are required to metamorphose the DRT-hierarchy of Montagovian logical subformulas into intensional profiles of (expositive) speech acts, on the one hand, and descriptions of its addressers' and addressees' current information states, on the other hand, which can then be readily compared with each other—in the form of some kind of "generalized truth-evaluation" (including the checking of all kinds of semantic and pragmatic felicity conditions), as is claimed in Section IV.

Then our software application ℜeALIS1.1 is demonstrated through discussing its different kinds of potential users and its main use cases for the users we call internal users and for those we call external users (Section V). Section VII presents the additional services of ℜeALIS2.1 as compared to ℜeALIS1.1 and an SDRT-based experimental software application called RUDI. We point out that ℜeALIS2.1 can be regarded as a model of the two parts of long-term memory—episodic and semantic memory—and this enables us to calculate new senses for words "in contexts hitherto unexperienced," as was formulated above (see also the illustration in Fig. 22).

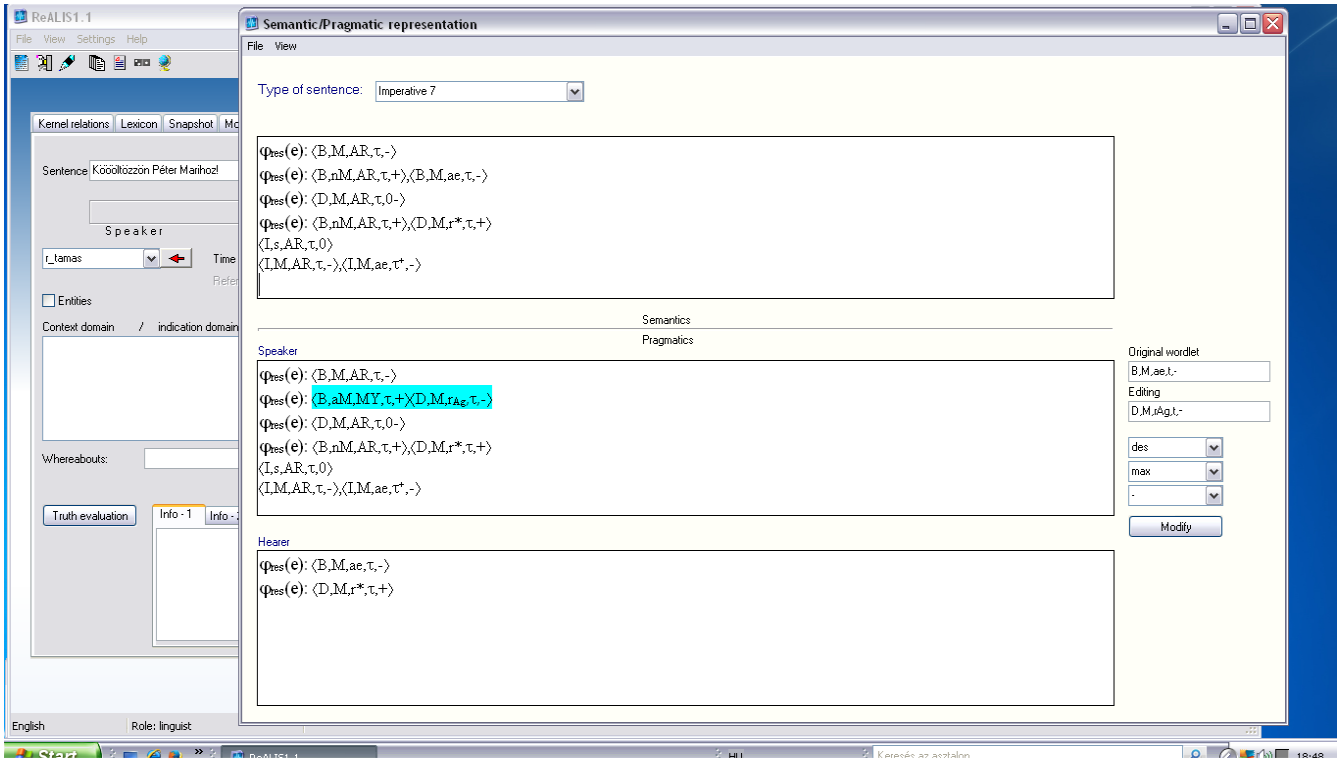


Figure 22. Comparison between the intensional profile of a proto-imperative speech act and the addresser's and the addressee's intensional profile.

ACKNOWLEDGMENT

We are grateful to SROP-4.2.2.C-11/1/KONV-2012-0005 (Well-Being in the Information Society). Special thanks are due to our colleagues Judit Kleiber, Veronika Szabó for their valuable comments on the pragmatic analyses presented in the paper.

REFERENCES

[1] G. Alberti and L. Nőthig, "ReALIS1.1: The Toolbox of Generalized Intensional Truth Evaluation," in Proc. of The Third International Conference on Intelligent Systems and Applications (INTELLI), J. Grzymala-Busse and I. Schwab, eds., 2014, pp. 60–66.

[2] G. Alberti, "ReALIS: An Interpretation System which is Reciprocal and Lifelong," Workshop 'Focus on Discourse and Context-Dependence' (16.09.2009, 13.30-14.30 UvA, Amsterdam Center for Language and Comm.), http://lingua.btk.pte.hu/realispapers.

[3] G. Alberti and M. Károly, "Multiple Level of Referents in Information State," LNCS7181, 2002, pp. 349–362.

[4] G. Alberti and J. Kleiber, "ReALIS: Discourse Representation with a Radically New Ontology," in Complex Visibles Out There. Olomouc Modern Language Series 4, L. Veselovská and M. Janebová, eds., Olomouc: Palacký University, 2014, pp. 513–528.

[5] G. Alberti, N. Vadász, and J. Kleiber, "Ideal and Deviant Interlocutors in a Formal Interpretation System," in The communication of certainty and uncertainty, A. Zuczkowski, R. Bongelli, I. Riccioni, and C. Canestrari, eds., Amsterdam: Benjamins, 2014, pp. 59–78.

[6] J. Kleiber and G. Alberti, "Uncertainty in Polar Questions and Certainty in Answers?," in Certainty-uncertainty – and the attitudinal space in between, Studies in Language Companion Series 165, S. Cantarini, W. Abraham, and E. Leiss, eds. Amsterdam: Benjamins, 2014, pp. 135–152.

[7] G. Alberti and J. Kleiber, "Where are Possibly Worlds? Arguments for ReALIS," Acta Linguistica Hungarica, vol. 59, 2012, pp. 3–26.

[8] G. Alberti, "Where are Possible Worlds? II. Pegs, DRSs, Worldlets and Reification," in Vonzásban és változásban, G. Alberti, J. Farkas, and J. Kleiber, eds. Pécs: Doctoral School of Linguistics at Univ. of Pécs, Hungary, 2012, pp. 308–323.

[9] G. Alberti, M. Károly, and J. Kleiber, "The ReALIS Model of Human Interpreters and Its Application in Computational Linguistics," in Software and Data Technologies: 5th International Conference, ICSOFT, J. Cordeiro, M. Virvou, and B. Shishkov, eds. Heidelberg: Springer, 2010, pp. 468–474.

[10] G. Alberti and M. Károly, "The Implemented Human Interpreter as a Database," in Proc. of Int. Conf. on Knowledge Engineering and Ontology Development, J. L. G. Diets, ed., Funchal: SciTePress, 2011, pp. 379–385.

[11] M. Károly and G. Alberti, "The Implementation of a ReALIS-based Method of Static Intensional Interpretation," in Proc. of 5th International Conference on Knowledge Engineering and Ontology Development, J. Filipe and J. G. L. Dietz, eds, 2013, pp. 393–398.

[12] I. Kilián and G. Alberti, "A Metamodel-Driven Architecture for Generating, Populating and Manipulating "Possible Worlds" to Answer Questions," in ICSOFT 2013 Proceedings, July 2013, pp. 74-78.

[13] J. Kleiber, "Across world(let)s in a representationist interpretation system," in Proc. of the 10th ESSLLI Student Session, J. Gervain, ed. 2005, pp. 112–121.

[14] Zs. Schnell and E. Varga, "Humour, Irony and Social Cognition," in Hungarian Humour. Humor and Culture 3, A. T. Litovkina, J. Szőllősy, P. Medgyes, and W. Chłopicki, eds., Cracow: Tertium Society for the Promotion of Language Studies, 2012, pp. 253–271.

[15] G. Alberti, M. Dóla, and J. Kleiber "Mood and modality in Hungarian: Discourse Representation Theory meets Cognitive Linguistics," Argumentum, vol. 10, Debrecen: Debrecen Univ. Press, Hungary, 2014, pp. 172–191.

[16] E. Varga, Zs. Schnell, T. Tényi, N. Németh, M. Simon, A. Hajnal, R. Horváth, E. Hamvas, S. Fekete, R. Herold, R. Járai, "Compensatory effect of general cognitive skills on non-literal language processing in schizophrenia," Journal of Neurolinguistics, vol. 29, 2014, pp. 1–16.

[17] G. Alberti, Zs. Schnell, and V. Szabó, "Autizmussal élni: más elmével élni [Living with Autism: Living with a Different Mind]," to appear. Budapest: L'Harmattan, 2015.

[18] G. Alberti and J. Kleiber, "The *GeLexi* MT Project," in Proceedings of EAMT 2004 Workshop (Malta), J. Hutchins, ed. Valletta: Univ. of Malta, 2004, pp. 1–10.

[19] G. Alberti and J. Kleiber, "The Grammar of ReALIS and the Implementation of its Dynamic Interpretation," Informatica (Slovenia) 34/2, 2010, pp. 103–110.

[20] K. Balogh and J. Kleiber, "Computational Benefits of a Totally Lexicalist Grammar," in Proc. of the 6th Conf. on Text, Speech and Dialogue, V. Matoušek and P. Mautner, eds. 2003, pp. 114–119.

[21] H. Kamp, J. van Genabith, and U. Reyle, "Discourse Representation Theory," in Handbook of Philosophical Logic, vol. 15, D. Gabbay and F. Guenthner, eds. Berlin: Springer, 2011, pp. 125–394.

[22] N. Asher and A. Lascarides, Logics of Conversation. Cambridge: Cambridge Univ. Press, 2003.

[23] J. L. Austin, How to Do Things with Words. Oxford: Clarendon Press, 1975 [1962].

[24] E. Oishi, "Discursive functions of evidentials and epistemic modals," in Certainty-uncertainty – and the attitudinal space in between, Studies in Language Companion Series 165, S. Cantarini, W. Abraham, and E. Leiss, eds. Amsterdam: Benjamins, 2014, pp. 239–262.

[25] D. R. Dowty, R. E. Wall, and S. Peters, Introduction to Montague Semantics. Dordrecht: Reidel, 1981.

[26] D. Schlangen, A. Lascarides, and A. Copestake, "Imperatives in dialogue," Perspectives on Dialogue in the New Millennium, Pragmatics & Beyond, New Series 114, P. Kühnlein, H. Rieser, and H. Zeevat, eds. Amsterdam: Benjamins, 2003, pp. 287–305.

[27] H. Kamp and U. Reyle, From Discourse to Logic. Dordrecht: Kluwer, 1993.

[28] J. van Eijck and H. Kamp, "Representing discourse in context," in Handbook of Logic and Language, J. van Benthem and A. ter Meulen, eds. Amsterdam: Elsevier, and Cambridge: MIT Press, 1997, pp. 179–237.

[29] N. Chomsky, Lectures on Government and Binding. Dprdrecht: Foris, 1981.

[30] N. Chomsky, Syntactic Structures. The Hague: Mouton, 1957.

[31] N. Chomsky, The Minimalist Program. Cambridge MA: MIT Press, 1995.

[32] K. É. Kiss and F. Kiefer, eds., The Syntactic Structure of Hungarian, Syntax and Semantics, vol. 27. New York: Academic Press, 1994.

[33] K. É. Kiss, The Syntax of Hungarian. Cambridge: Cambridge Univ. Press, 2002.

[34] G. Alberti, "Restrictions on the Degree of Referentiality of Arguments in Hungarian Sentences," Acta Lingiustica Hungarica, vol. 44(3-4) , 1997, pp. 341–362.

[35] F. Heck, G. Müller, R. Vogel, S. Fischer, S. Vikner, and T. Schmid, "On the nature of the input in optimality theory," The Linguistic Review, vol. 19, 2002, pp. 345–376.

[36] J. L. Austin, "Performative Utterances," in Philosophical Papers, J. O. Urmson and G. J. Warnock, Oxford: Oxford University Press, 1979 [First edition 1961], pp. 233–252.

[37] K. Bach and R. M Harnish, Linguistic Communication and Speech Acts. Cambridge, MA: MIT Press, 1979.

[38] P. Labinaz and M. Sbisà, "Certainty and uncertainty in assertive speech acts," in The communication of certainty and uncertainty, A. Zuczkowski, R. Bongelli, I. Riccioni, and C. Canestrari, eds., Amsterdam: Benjamins, 2014, pp. 31–58.

[39] P. Pelyvás, "Relating Cognitive Models In Nominal Grounding. Referential, Attributive Use and Referential Opacity: A Case for a Blend?" in Metaphors of sixty. Papers presented on the occasion of the 60th birthday of Zoltán Kövecses, R. Benczés and Sz. Csábi, eds. Budapest: SEAS, ELTE, 2006, pp. 196–209.

[40] J. Nuyt, "Subjectivity in modality, and beyond," in The communication of certainty and uncertainty, A. Zuczkowski, R. Bongelli, I. Riccioni, and C. Canestrari, eds., Amsterdam: Benjamins, 2014, pp. 13–30.

[41] J. Groenendijk, M. Stokhof, and F. Veltman, "Coreference and Modality," in The Handbook of Contemporary Semantic Theory, Sh. Lappin, ed. Oxford: Blackwell, 1996, pp. 179–213.

[42] N. Marsali, "Lying as a scalar phenomenon: Insincerity along the certainty-uncertainty continuum," in Certainty-uncertainty – and the attitudinal space in between, Studies in Language Companion Series 165, S. Cantarini, W. Abraham, and E. Leiss, eds. Amsterdam: Benjamins, 2014, pp. 239–262.

[43] F. Nissim and R. Dyckhoff, "Proof-theoretic semantics for a natural language fragment," Linguistics and Philosophy, vol. 33 (6), 2010, pp. 447–477.

[44] C. Pollard, "Hyperintensions." ESSLLI 2007, http://www.cs.tcd.ie/esslli2007, 2007.

[45] K. E. Judge, "Epistemic uncertainty and the syntax of speech acts," in Certainty-uncertainty – and the attitudinal space in between, Studies in Language Companion Series 165, S. Cantarini, W. Abraham, and E. Leiss, eds. Amsterdam: Benjamins, 2014, pp. 217–237.

[46] A. Kratzer, "What 'must' and 'can' must and can mean," Linguistics and Philosophy, vol. 1, pp. 337–355, 1977.

[47] A. Kratzer, Modals and Conditionals. Oxford: OUP, 2012.

[48] R. W. Langacker, "Remarks on Nominal Grounding," Functions of Language 11:1. 2004, pp. 77–113.

[49] G. Alberti, "Climbing for Aspect – with no Rucksack," In Verb Clusters; A study of Hungarian, German and Dutch. Linguistics Today, vol. 69, K. É. Kiss and H. van Riemsdijk, eds. Amsterdam/Philadelphia: Benjamins, pp. 253–289, 2004.

[50] E. Leiss, "Modes of modality in an Un-Cartesian framework," in Certainty-uncertainty – and the attitudinal space in between, Studies in Language Companion Series 165, S. Cantarini, W. Abraham, and E. Leiss, eds. Amsterdam: Benjamins, 2014, pp. 47–62.

[51] Z. Bódis, J. Kleiber, É. Szilágyi, and A. Viszket, "LiLe projekt: adatbázis mint "dinamikus korpusz" [The 'Linguistic Lexicon' Project: data base as a dynamic corpus]," MSzNy2 [Second Conference on Hungarian Computational Linguistics], Z. Alexin and D. Csendes, eds. Szeged: Juhász Publ., pp. 11–18.

[52] J. Kleiber, "Total Lexicalism in Language Technology," in Proc. of the 12th ESSLLI Student Session, V. V. Nurmi and D. Sustretov, eds. 2007, pp. 149–160.

[53] G. Alberti and I. Kilián "Bipolar Influence-Chain Families Instead of Lists of Argument Frames," in Well-Being in Information Society Conference Proceedings, G. Rappai and Cs. Filó, eds. November 2015, Pécs: Univ. of Pécs, pp. 4-16.

[54] I. Kilián and G. Alberti "Bipolar Influence-Chain Families Instead of Lists of Argument Frames, II. Cornerstones of Implementation in Prolog" in Well-Being in Information Society Conference Proceedings, G. Rappai and Cs. Filó, eds. November 2015, Pécs: Univ. of Pécs, pp. 17-23.

[55] K. Balogh and J. Kleiber, "A Morphology Driven Parser for Hungarian," in Proc. of the 5th Tbilisi Symposium on Language, Logic and Computation, R. Asatiani, K. Balogh, G. Chikoidze, P. Dekker, and D. de Jongh, eds. Amsterdam/Tbilisi: ILLC, University of Amsterdam / CLLS, Tbilisi State University, 2004, pp. 29–37.

[56] É. Szilágyi, "The Rank(s) of a Totally Lexicalist Syntax," in Proc. of the 13th ESSLLI Student Session, K. Balogh, ed. 2008, pp. 175–184.

[57] G. Alberti, M. Károly, and J. Kleiber, "From Sentences to Scope Relations and Backward," in Natural Language Processing and Cognitive Science, Proceedings of NLPCS 2010, B. Sharp and M. Zock, eds. Funchal: SciTePress, 2010, pp. 100-111.

[58] G. Alberti, J. Kleiber, and A. Viszket, "GeLexi Project: Sentence Parsing Based on a GEnerative LEXIcon," Acta Cybernetica 16 (Hungary). 2004, pp. 587–600.

[59] G. Alberti, K. Balogh, J. Kleiber, and A. Viszket, "Total Lexicalism and GASGrammars: A Direct Way to Semantics," LNCS 2588, Berlin: Springer, 2003, pp. 37–48.

[60] L. Nőthig and G. Alberti "The Discourse-Semantic and Syntactic Background Behind ℜeALIS," in Well-Being in Information Society Conference Proceedings, G. Rappai and Cs. Filó, eds. November 2015, Pécs: Univ. of Pécs, pp. 104–129.

[61] J. Seligman and L. S. Moss, "Situation Theory," in Handbook of Logic and Language, J. van Benthem and A. ter Meulen, eds. Amsterdam: Elsevier, and Cambridge: MIT Press, 1997, pp. 239–309.

[62] F. Landman, Towards a Theory of Information. Dordrecht: Foris, 1986.

[63] U. Reyle, "Dealing with Ambiguities by Underspecification: Construction, Representation and Deduction," Journal of Semantics, vol. 10, 1993, pp. 123–179.

[64] A. Szabolcsi, ed., Ways of Scope Taking, SLAP 65. Dordrecht: Kluwer, 1997.

[65] J. Farkas and M. Ohnmacht, "Aspect and Eventuality Structure in a Representational Dynamic Semantics," in Vonzásban és változásban, G. Alberti, J. Farkas, and J. Kleiber, eds. Pécs: Doctoral School of Linguistics at Univ. of Pécs, Hungary, 2012, pp. 353–379.

[66] G. Leech, Principles of Pragmatics. Harlow: Longman, 1983.

[67] B. Gyuris, "Megjegyzések a pragmatika tárgyáról és hasznáról [Comments on the subject and the benefit of pragmatics]," Magyar Nyelv [Hungarian Language], vol. 109 (2), 2013, pp. 162–170.

[68] W. C. Mann, "Models of Intentions in Language," Perspectives on Dialogue in the New Millennium, Pragmatics & Beyond, New Series 114, P. Kühnlein, H. Rieser, and H. Zeevat, eds. Amsterdam: Benjamins, 2003, pp. 166–178.

[69] A. Zuczkowski, R. Bongelli, L. Vincze, and I. Riccioni, "Epistemic stance," in The communication of certainty and uncertainty, A. Zuczkowski, R. Bongelli, I. Riccioni, and C. Canestrari, eds., Amsterdam: Benjamins, 2014, pp. 115–135.

[70] P. Grice, "Logic and Conversation," in Speech Acts, Syntax and Semantics, vol. 3., P. Cole and J. L. Morgan, eds. New York: Academic Press, 1975, pp. 41–58.

[71] H.-M. Gärtner and B. Gyuris, "Pragmatic markers in Hungarian: Some introductory remarks," Acta Linguistica Hungarica, vol. 59, 2012, pp. 387–426.

[72] A. Schirm, A diskurzusjelölők funkciói: A *hát*, az *-e* és a *vajon* elemek története és jelenkori szinkrón státusa alapján [Functions of discourse markers: on the basis of the history and the synchronic status of *hát*, *-e* and *vajon*]. PhD dissertation, Univ. of Szeged, 2011.

[73] Ch. Potts, "Conventional Implicatures, a Distinguished Class of Meanings," in The Oxford Handbook of Linguistic Interfaces, G. Ramchand and Ch. Reiss, eds. Oxford: Oxford University Press, 2007, pp. 475–501.

[74] Ch. Potts, "Presupposition and Implicature," to appear in The Handbook of Contemporary Semantic Theory, Sh. Lappin and Ch. Fox, eds. Oxford: Wiley-Blackwell, 2013.

[75] J. R. Searl, Speech Acts: An Essay in the Philosophy of Language. Cambridge, MA: Cambridge University Press, 1969.

[76] J. R. Searl, Expression and Meaning. Cambridge, MA: Cambridge University Press, 1979.

[77] H.-M. Sohn, Korean. New York: Routledge, 1994.

[78] A. Y. Aikhenvald, Imperatives and Commands. New York: Oxford University Press, 2010.

[79] N. Asher and A. Lascarides, "Imperatives in dialogue," Perspectives on Dialogue in the New Millennium, Pragmatics & Beyond, New Series 114, P. Kühnlein, H. Rieser, and H. Zeevat, eds. Amsterdam: Benjamins, 2003, pp. 1–24.

[80] I. Vaskó, "Pragmatic particles indicating expectation – The case of *persze*," Acta Linguistica Hungarica, vol. 59, 2012, 465–486.

[81] J. Schrickx, "Latin commitment-markes: *scilicet* and *videlicet*," in Certainty-uncertainty – and the attitudinal space in between, Studies in Language Companion Series 165, S. Cantarini, W. Abraham, and E. Leiss, eds. Amsterdam: Benjamins, 2014, pp. 285–296.

[82] M. Szücs, "A *hadd* problémaköre [The problem of *hadd* 'let']," LingDok, vol. 9, Szeged: Univ. of Szeged, 2010, pp. 193–210.

[83] A. Péteri, "The Hungarian Imperative Particle *Hadd*," Acta Linguistica Hungarica, vol. 59, 2012, pp. 439–463.

[84] G. Turi, "Kötőmód a mai magyar nyelvben [The subjunctive in present-day Hungarian]," Argumentum, vol. 5., 2009, pp. 25–38.

[85] R. W. Langacker, "Cognitive Grammar," in The Oxford Handbook of Cognitive Linguistics. Oxford: Oxford Univ. Press, 2007, pp. 421–462.

[86] D. R. Dowty, Word Meaning and Montague Grammar. Dordrecht: Reidel, 1979.

[87] I. Kilián "ℜeALkb. Towards of a semantic platform," in Well-Being in Information Society Conference Proceedings, G. Rappai and Cs. Filó, eds. November 2015, Pécs: Univ. of Pécs, pp. 130–136.

[88] J. Farkas, A finn nyelv (indexelt) generatív szintaxisa [An (indexed) generative syntax of Finnish]. Pécs PhD Theses, vol. 2, I. Kassai, ed. Pécs: Doctoral School of Linguistics, 2009.

[89] K. Lochbaum, "A Collaborative Planning Model of Intentional Structure," Computational Linguistics, vol. 24(4), 1998, 525–572.

[90] E. Tulving, "Episodic Memory and Autonoesis. Uniquely Human?" in The Missing Link in Cognition: Origins of Self-Reflective Consciousness, H. S. Terrace and J. Metcalfe, eds. Oxford: Oxford Univ. Press, 2005, 3–56.

[91] J. Pustejovsky, The Generative Lexicon. Cambridge, Mass. & London: MIT Press, 1995.

[92] M. Károly "Possibilities of refining the ℜeALIS world and language model and extending its semantic postulate set – from the scope of cognitive and functional linguistics, illustrated by interpretations of some static Hungarian verbs," in Well-Being in Information Society Conference Proceedings, G. Rappai and Cs. Filó, eds. November 2015, Pécs: Univ. of Pécs, pp. 137-146.

[93] F. Kiefer, Jelentéselmélet [Theory of Meaning]. Budapest: Corvina, 2000.

Binding of Security Credentials to a specific Environment on the Example of Energy Automation

Rainer Falk and Steffen Fries
Corporate Technology
Siemens AG
Munich, Germany
e-mail: {rainer.falk|steffen.fries}@siemens.com

***Abstract*—Information security in critical infrastructures is becoming an inevitable part of networked control systems. Examples are industrial automation, process automation, and energy automation systems. Characteristic for all these systems is the data exchange between intelligent electronic devices – IEDs, which are used to monitor and control the operation. In energy automation deployments these IEDs provide the data for a obtaining a system view of connected energy resources. This becomes increasingly important as the number of decentralized energy resources – DER – is constantly increasing. Based on the system view, a set of DER, building a virtual power plant, can be managed reliably. The communication is realized through domain-specific communication protocols like IEC 61850, or IEC 60870-5. This communication is performed over networks of different administrative domains, also over public networks. Therefore, IT security is a necessary prerequisite to prevent intentional manipulations, thereby supporting the reliable operation of the energy grid. Basis for protecting metering and control communication are cryptographic security credentials, which need to be managed not only during operation, but most importantly during installation (initial enrollment). This process needs to be as simple as possible to not increase the overall effort and to not introduce additional sources for failures. Hence, automatic credential management is needed to ensure an efficient management for a huge number of devices. This paper describes a new approach for the automatic initial security credential enrollment process during the installation phase of IEDs. The approach targets the binding of the security credentials of the installed IEDs to the operational environment and also to the intended utilization of the IED by embedding specific information into the enrollment communication, which is then reflected in the issued X.509 certificates.**

Keywords–security; device authentication; automated certificate enrollment; real-time; network access authentication; firewall; substation automation; smart grid; smart energy; DER; PKI; IEC 61850; IEC 60870-5; IEC 62351

I. INTRODUCTION

Decentralized energy generation, e.g., through renewable energy sources like solar cells, or wind power, is becoming increasingly important to generate environmentally sustainable energy, and thus to reduce greenhouse gases leading to global warming. Introducing decentralized energy generators into the current energy distribution network poses great challenges for energy automation as decentralized energy generation needs to be monitored, and controlled to a similar extend as centralized energy generation in power plants. This requires widely distributed communication networks. Distributed energy generators may also be aggregated on a higher level to form a so-called virtual power plant (VPP). Such a virtual power plant may be viewed from the outside in a similar way as a common power plant with respect to energy generation capacity. But due to its decentralized nature, the demands on communication necessary to control the virtual power plant are much more challenging. Moreover, these decentralized energy resources may also be used in an autonomous island mode, without any connection to a backend system.

Furthermore, the introduction of controllable loads on residential level requires enhancements to the energy automation communication infrastructure as used today. Clearly, secure communication between a control station, and equipment of users (e.g., decentralized energy generators) as well as with decentralized field equipment must be addressed. Standard communication technologies as Ethernet and the Internet protocol IP are increasingly used in energy automation environments down to the field level [1] [2][3].

Figure 1 depicts smart energy automation scenarios showing the increased communication demand, e.g., through the integration of microgrids, controllable loads, and also electro mobility. IT security is a base requirement to be addressed in all the scenarios to ensure the reliable operation of the smart grid. The communication in energy automation systems must therefore be cryptographically secured. For energy automation control protocols like IEC 61850 [4], or IEC 60870, the accompanying standard series IEC 62351 [6] is available defining various options to secure communicated data. One base for secure interaction are security credentials in the form of X.509 digital certificates, corresponding private keys, and a related security policy. All need to be provisioned during device installation, and maintained during operation. Especially, the exchange of devices with spare parts should not lead to breaches in security, which could occur if the key material of the replaced devices is not handled appropriately. This is especially important as IEDs may not always be placed in a physically protected environment, and even publicly accessible.

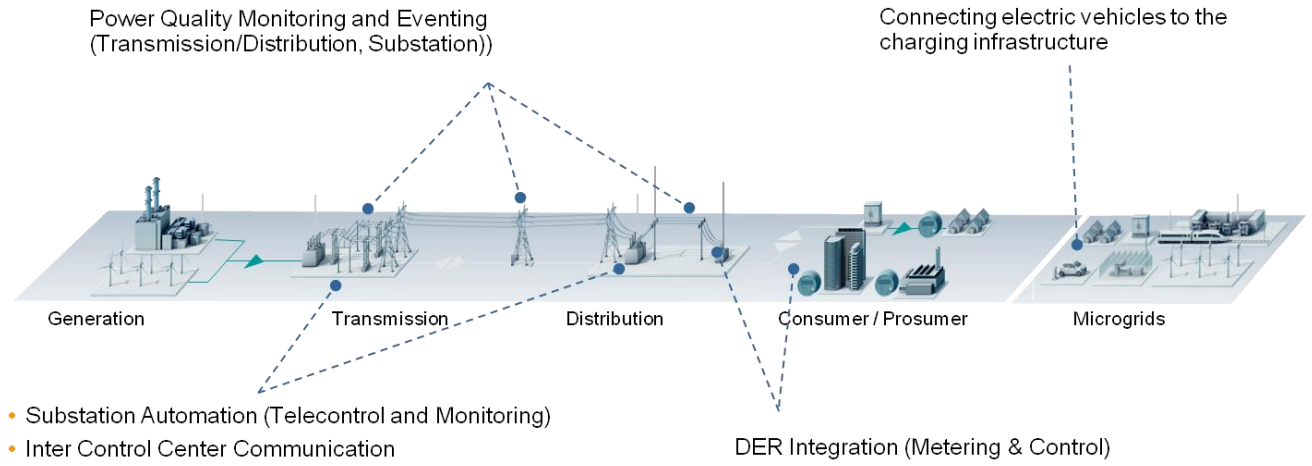


Figure 1. Typical Smart Energy Automation Scenarios

To ensure that this key material cannot be misused, e.g., in the context of an unintended service, or in an unintended environment, the key material has to be bound to the respective device and/or application purpose. Existing options, e.g., using key usage extensions in X.509 certificates, may not always be sufficient, as they relate to the actual usage of the cryptographic key, and not to the device application environment. The purpose-binding of a cryptographic key described in this paper therefore restricts the key acceptance depending on location information, and potential other parameters.

The remainder of this paper is structured as follows: Section II provides an overview of two example Smart Grid use cases. Section III depicts an overview of secure communication with respect to the use cases explained before. This section motivates the handling of security key material. Section IV provides an overview about the utilized security credentials, while Section V introduces the Public Key Infrastructure as means for credential handling. Section VI introduces existing certificate enrollment methods, while Section VII describes an enhancement to have purpose bound certificates. Section 0 concludes the paper and provides an outlook.

II. SMART ENERGY AUTOMATION USE CASES

To motivate communication security, two example use cases are addressed in this paper, substation automation, and DER incorporation in energy control networks. They are explained in the following two subsections.

A. Substation Automation

Automation networks are typically shared networks connected in a ring, star, or bus topology, or a mixture of these. Most often, the time-critical part is realized on a dedicated network segment, while the rest of the communication supporting the automation systems is performed on networks with lower performance requirements.

An example for energy automation is the communication within a substation. A substation typically transforms voltage levels, and includes power monitoring, and protection functions. The example in Figure 2 shows a typical setup of a primary substation. The red rectangle shows the area, in which the IEDs communicate status information, and provide this information into the substation automation zone, and further up the hierarchy to the control center.

As depicted in Figure 2, the substation bus can be realized as communication ring, connecting the protection relays, acting in real-time typically via Ethernet. Ideally, the network is separated to already provide a first security barrier with different access restrictions. There is a connection to other zones within the substation, separated from the real-time part using Firewalls. Examples are the automation zone, or the remote access zone. Another example is the zone storing the historian information also interacting with a backend SCADA system. The historian server is a device for archiving measurements, events, and alarms of the substation.

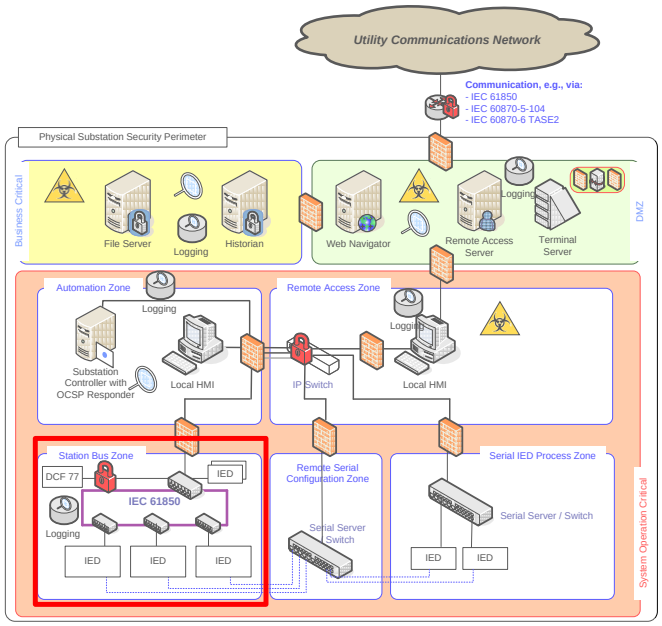


Figure 2. Substation – Functional Split into Zones

Figure 2 above already shows security elements deployed within a substation, like Firewalls, virus checking tools, or access control means to components, or data as recommended in IEC 62351-10 [6].

B. DER Incorporation

Decentralized Energy Resources (DER) may be connected to the Smart Grid at two different connection points. Depending on the amount of energy provided, they may be connected to the low voltage network, or to the medium voltage network (distribution network). The first one is rather typical for DER in residential areas, like a solar panel, while the connection to the medium voltage network is done for larger deployments like wind power farms, or solar parks. Necessary for both is the connectivity to a communication infrastructure to allow a control center to act on provided information about current energy generation, but also to provide scheduling information to the DER, e.g., depending on the weather forecast, to better balance the feed in of energy into the electrical network. Communication with the DER may be done using different communication technologies, like Power Line Communication (PLC), or wireless communication via the UMTS network.

Figure 3 shows an example integration of DER utilizing a mapping of IEC 61850 to the XMPP (eXtensible Message and Presence Protocol, [7]), which is currently being discussed within the standardization. This approach allows connecting to DERs, which reside in a customer's network behind a firewall. For the distribution network operator (DNO), it is essential to know, which DERs are associated to his operational control.

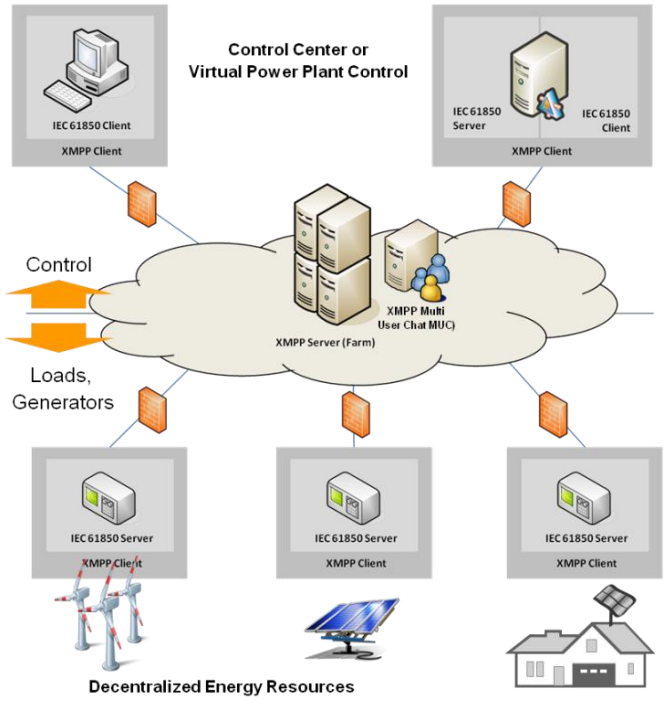


Figure 3. DER Integration based on IEC 61850 over XMPP

This can be supported by the used security credentials using additional information depending, e.g., on the geographic location of a DER, or on the association with a dedicated DNO.

III. Secure Communication in Smart Grid

IEC 61850 [4], [5] is a standard for communication in the domain of energy automation. It is envisaged to be the successor of the currently used standards IEC 60870-5-104, and DNP3 especially used in the North American region. IEC 61850 enables interoperability between devices used in energy automation. For example, two IEC 61850 enabled devices of different vendors can exchange a set of clearly defined data, and the devices can interpret, and use these data to achieve the functionality required by the application due to a standardized data model. This is facilitated by the IEC 61850 series specifying:

- an Abstract Communication Service Interface (ACSI),
- a semantic model based on an object oriented architecture,
- specific Communication Service Mappings (SCSM),
- a project engineering workflow including a configuration description language (SCL) based on the XML language.

In particular, IEC 61850 enables continuous communication from a control station to decentralized energy generators, or to IEDs (like protection relays) in a substation.

IT security is increasingly important in energy automation as on part of the Smart Grid. Here, the IEC 62351 framework [6], shown in Figure 4, with currently 13 defined parts kicks in, defining security services for IEC 61850 based communication covering different deployment scenarios using serial communication, IP-based communication, and also Ethernet communication. The latter one is used locally within a substation to cope with the high real-time requirements. While it may be not always necessary to encrypt the communication to protect confidentiality, there is a high demand to protect the communication against manipulation, and to allow for source authentication.

IEC 62351 relies on existing security technologies as much as possible, and profiles it for the application environment. One example is the application of Transport Layer Security (TLS, RFC 5246 [8]) to protect TCP-based communication. Here, IEC 62351-3 basically reduces the manifold options of TLS to ease interoperability. Another example is the adoption of Group Domain of Interpretation (GDOI, RFC 6407 [9]) as group-based key management in IEC 62351-9 to distribute key material for the protection of status information, and event signaling between IEDs in a substation, or across substations using Wide Area Networks (WANs).

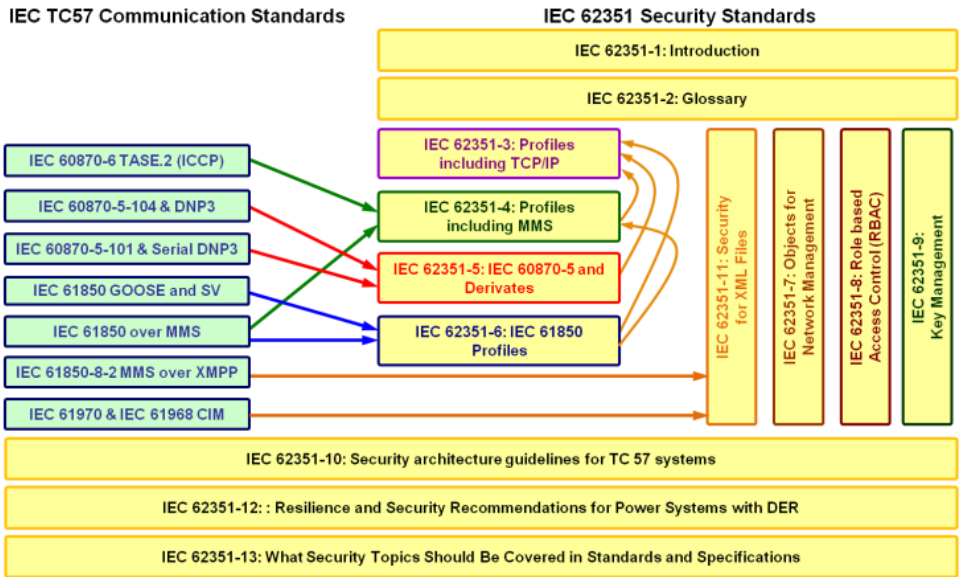


Figure 4. IEC 62351 series relation to energy automation protocols

A specific characteristic throughout IEC 62351 is the consequent application of X.509 certificates, and corresponding private keys for mutual authentication on network layer, and application layer. This requires an efficient handling of X.509 key material, and the availability of this information right from the installation. There is a strong need to provide these credentials without increasing the installation effort. For instance, devices may generate their own key pair, but certification needs to bind this key pair to the operational. This is a challenge from the pure technical perspective as a high number of devices need to be equipped with the key material. But also from the network operator process point of view this is challenging, as the key material has a lifecycle, and needs to be updated once in a while. These aspects will be addressed in the following sections.

IV. SECURITY CREDENTIALS

Security credentials are used for different purposes, and in different phases of the communication lifetime, and may comprise:

- Certificates and corresponding private keys: Used to authenticate entities and support session key establishment.
- Pre-shared keys (e.g., for real-time communication): Used as a shared secret between entities to build up mutual authentication between them. This scheme may be used when lacking a PKI, or when doing an initial authentication. In the latter, it may be an entity password, which allows an entity to authenticate itself against the Certificate Authority (CA), for example when performing a certificate signature request (CSR).
- Session secret keys (pair wise or group-based): Used for efficient encryption, or integrity checks on communication messages.
- Session parameters (dedicated cryptographic algorithms): Used to specify the lifetime of the session keys.
- Cryptographic access tokens: Mostly used to transfer/provide authorization/access of resources to entities for a limited time.

For the context of this document, the focus lies on X.509 certificates, and corresponding private keys.

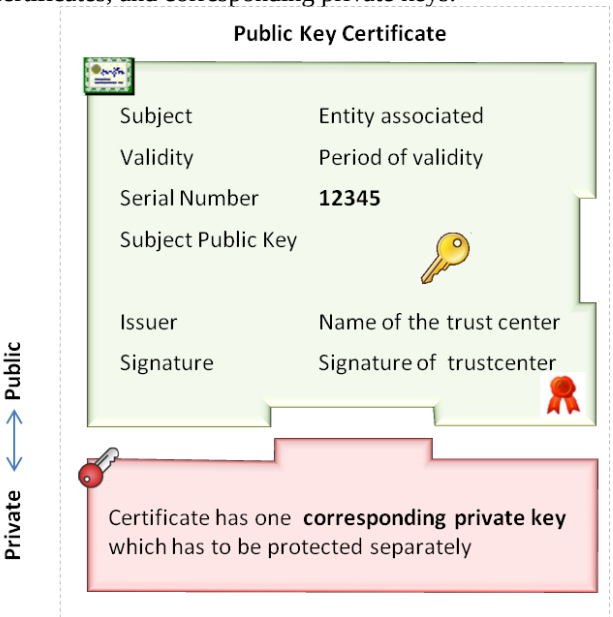


Figure 5. Structure of a X.509 public-key credential

According to X.509 [10], and as shown in Figure 6, a public-key certificate is an electronic document that binds the identity of entity to the public private key pair of that entity. This binding is verified by a digital signature of the issuing CA building the certificate. In addition to the public-key, and identity of the owner of public-key certificate, the

public-key certificates holds verified information about validity period, and the identity of the issuer.

A public-key certificate may include extensions providing additional information. An extension is identified by an object identifier allocated by the organization defining the extension. The use of X.509 attribute certificates, shown in Figure 6, can be an effective way to separate the management of identity from the management of authorizations associated with an identity. An attribute certificate, as shown in Figure 6, provides an option for temporary enhancement of public key certificate, linked with the public-key certificate's unique identifier.

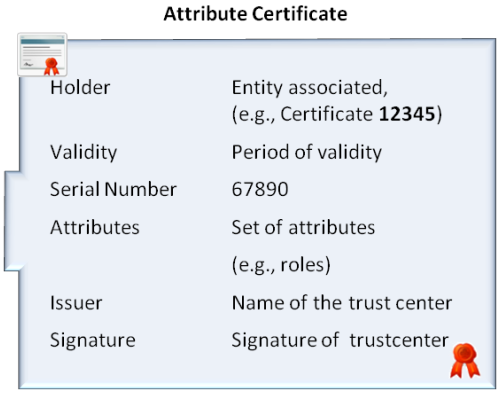


Figure 6. X.509 Attribute Certificate structure

The combination of public key certificates, and attribute certificates provides high flexibility for operational use cases. One reference can be given by IEC 62351-8 [6] utilizing this approach for supporting Role-based Access Control. A different approach is provided by FlexiCert [11], allowing the inclusion of additional attributes during the lifetime of a public key certificate.

For the context of this paper, the focus is placed on X.509 public key certificates, which are issued for IEDs. The following section provides more details about the handling of these credentials using a PKI.

V. PUBLIC KEY INFRASTRUCTURE - PKI

A PKI typically contains a variety of services requiring interfaces in the devices utilizing the PKI, and also an accompanying process. In general, a PKI provides a secure, reliable, and scalable environment for the complete lifecycle of key material, i.e., generating, distributing, and querying public keys for secrecy, correctness, and sender verification. Moreover, it binds the "owner" to the public key using a digital certificate, and thus enables identification of users, and components utilizing certificates. Furthermore, it maintains and distributes status information for the lifetime of that binding, i.e., from the generation till the revocation. The general functionality and formats are described in RFC 5280 [10].

The following list provides a short overview about the different components, which are depicted in Figure 7Figure 7.

- **Registration authority (RA)** authenticates the user, or IED, or the data submitted by the user, or IED, performs an authorization check, and initiates the certificate generation at the CA. For machine-to machine communication, the RA can be used to mediate between the device applying for a certificate, and the CA.
- **Certification authority (CA)** is a trusted entity that certifies public-keys by issuing certificates.
- **Key/certificate archive** is a repository in which the CA stores certificates and/or generated key pairs.
- **Key generation** is a function of the PKI responsible for the generation of key material (public, and private keys), which are certified through the CA.
- **Public Directory** is a (usually publicly readable) database to which the CA stores all issued certificates.
- **Revocation Lists** are also a publicly readable database to which the CA stores all revoked certificates.

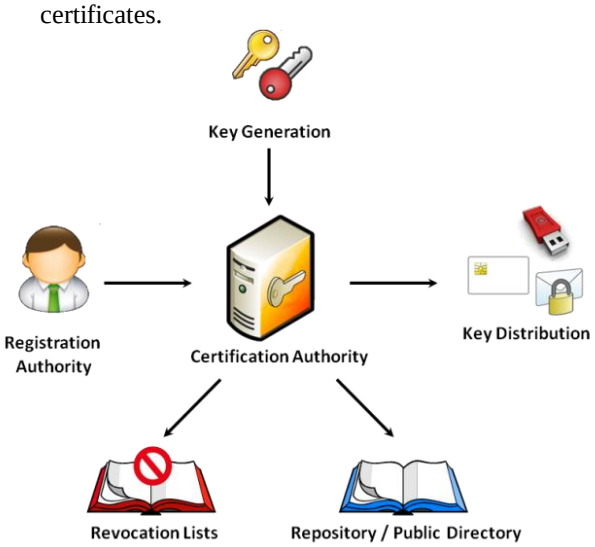


Figure 7. PKI Components

In the context of smart grid, a PKI may be operated by a utility company as internal PKI, or it may be a public PKI, also depending on the target use case, and the need for interoperation between different parties. Moreover, the functionality provided by the PKI needs to be streamlined to the target environment to avoid unnecessary effort. In any case, the devices utilizing key material issued by the CA need to provide the technical interfaces to accomplish this task. This is described in more detail in IEC 62351-9 targeting the key management explicitly.

Section VII describes an enhancement of the typical used PKI setup by introducing an intermediary, which provides all operational environment specific information. This avoids the pre-configuration of IEDs with this information.

VI. EXISTING CERTIFICATE ENROLLMENT METHODS

This section describes common methods for certificate enrollment taking device capabilities into account. Capabilities in this context relate to local, and remote key generation. Typically, local generation of key material is desired to avoid the handling of private keys outside the devices. Note that depending on the key usage, there may be requirements to also have the private key available in a trust

center to ensure that encrypted information can be accessed even the device hosting the private key was either damaged, or has been compromised.

A. Manual Enrollment

Manual enrollment relates to the manual connection of a device to an engineering tool to provide the key material during a local configuration session, prior to the connection in the target network. This approach requires a significant initial configuration effort, and is especially cumbersome in case of device replacements.

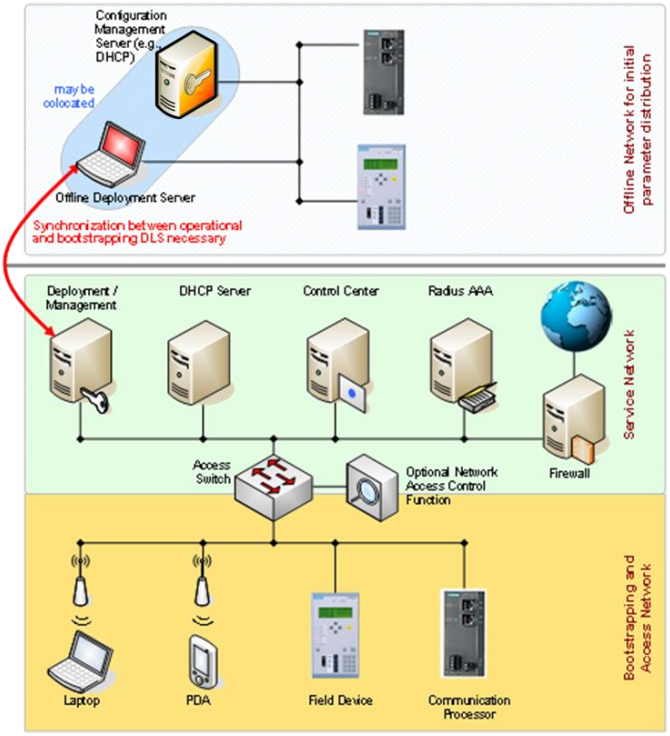


Figure 8. Manual enrolment using offline engineering

It may be realized using an offline engineering network to bootstrap the security credentials for connected devices, as shown in Figure 8. Here, the devices, or components may not possess a cryptographic credential up front, as the separate network is assumed to be physically secure. In the simplest form, it may be a direct connection of an engineering laptop to the component to be administered using purely local point-to-point communication.

B. Automated Enrollment

Automated enrollment refers to the initial configuration of devices including the key material. This is shown in Figure 9. Field devices are connected to the network, and contact the PKI server to obtain certified key material. Here, the field devices generate their public/private key pairs locally, and send a Certificate Signing Request (CSR) for the public key to the RA/CA (part of the PKI server). Part of the CSR may be a serial number of the device, against which the PKI server can check a configured list of devices allowed to be enrolled. This authorization may also be realized by other means like one-time passwords.

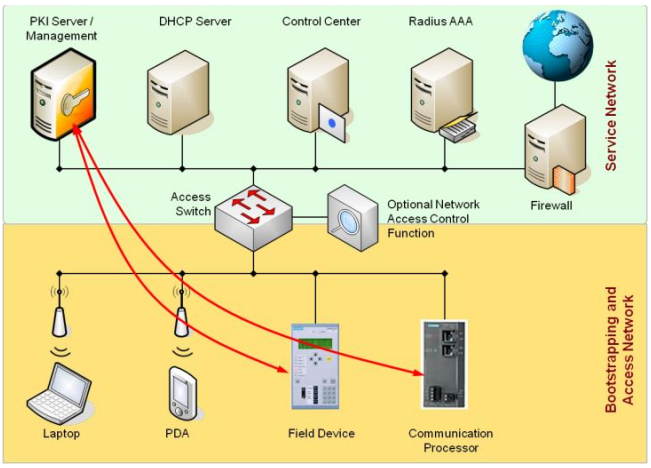


Figure 9. Automated distribution using management protocols

According to RFC 2986 [12] the CSR is defined in ASN.1 as shown in Figure 10 below.

```
CertificationRequest ::= SEQUENCE {
   certificationRequestInfo
     CertificationRequestInfo,
   signatureAlgorithm AlgorithmIdentifier{{
     SignatureAlgorithms }},
   Signature BIT STRING
   }

CertificationRequestInfo:
   CertificationRequestInfo ::= SEQUENCE {
   version       INTEGER { v1(0) } (v1,...),
   subject       Name,
   subjectPKInfo SubjectPublicKeyInfo{{
    PKInfoAlgorithms }},
   Attributes [0] Attributes{{
    CRIAttributes }}
   }

SubjectPublicKeyInfo { ALGORITHM : IOSet}
    ::= SEQUENCE {
    Algorithm AlgorithmIdentifier {{IOSet}},
    subjectPublicKey BIT STRING
   }
```

Figure 10. Certification Request structure [12]

As shown in Figure 10, a CSR is a self-contained structure providing the necessary information for the certificate application and issuing process but also a protection of this information regarding its authenticity. This is done through a signature with the private key corresponding to the public key in the CSR itself. This signature provides a proof of possession of the private key on at the requesting entity towards the issuing CA.

Several protocols are known for transmitting a CSR to a CA. Examples are:

- SCEP – Simple Certificate Enrollment Protocol [13] was developed to simplify the enrollment of large numbers of devices, and to make issuing, and revocation of digital certificates as scalable as possible. Entities can use SCEP to request their digital certificate electronically using the PKI

certificate forms PKCS#7 and PKCS#10 over HTTP. Note that SCEP cannot be used in conjunction with RAs using elliptic curve based certificates.

- CMP – Certificate Management Protocol [14] was developed using an own syntax to transport PKCS#10 container. It defines a protocol for the interaction between a client, and the PKI components. It provides more options compared to SCEP but is also more complex. Besides CRL retrieval, certification request handling, identification/authorization option for the requester as well as proof of possession of the associated private key CMP provides additional functionality like cross certification, and certificate revocation, which is to be supported mandatory. CMP supports the client side, and server side generation of key material.
- CMC – Certificate Management over CMC [15], utilizes PKCS#7, and PKCS#10. It provides more mandatory options compared to SCEP, and results therefore in a higher complexity. Besides CRL retrieval, certification request handling, identification, and authorization options for the requester, as well as proof of possession of the associated private key, CMC provides additional functionality like cross certification, and certificate revocation. CMC defines a simple, and a full PKI request/response handshake, but requires both to be implemented. CMC supports the client, and server side generation of key material.
- EST – Enrollment over Secure Transport [16] is discussed below more elaborately, to provide an overview about the general information exchange, and setup for enrollment.
- XML Key Management Specification [17]

In general, these protocols describe the communication of a CSR from a device to the CA, were the device ideally generates the key pair for itself. Additionally to identification information like the serial number, further information can be connected with the CSR, like a password (to be used to authorize a potential future revocation), or key usage restrictions. The CSR has to be protected to prevent illegitimate issuing of certificates. The CSR itself may be protected using the public key of the RA/CA as in case of SCEP. In case of CMP, the CSR is protected using an initial authentication key, and in EST, the CSR is transmitted over a secured communication link. Here TLS is applied, providing the opportunity to authenticate both peers during the connection establishment. Also, there may be an intermediate RA located between the device sending the CSR, and the CA, which already performs the verification of the CSR to reduce the load on the CA.

In Figure 11, EST is taken as example to provide more insight to the communication during an enrollment session. EST bases on CMC, and defines some of the CMC functionality as optional resulting in reduced complexity. Here, only the simple PKI request/response interaction is mandatory, while the full procedure support is optionally. From a functionality perspective EST can be seen as evolvement of SCEP. EST utilizes TLS as secure channel, and leverages the authentication of the TLS channel for identification, and authorization of the requester by binding the CSR to the actual TLS session.

The CMC part provides proof-of-possession of the private key corresponding to the public key in the CSR. Besides the CSR processing, EST allows for exchange of CA certificates, and corresponding chains, as well as for renewal. Moreover, it supports certificate attribute retrievals from a client side to query additional information, or boundary conditions prior to generating a CSR. EST supports the distribution of the operational CA's certificates in a "enrollment preparation phase". This distribution already needs to be secured as the CA's root certificate constitutes the trust anchor for the operational environment. According to EST, this exchange may be secured using initially available credentials, which may be a vendor certificate, or a shared secret available on the device. From an automation perspective the device vendor's certificate is preferred, as it allows the production of the devices as well as the central administration of the available credential in the operational environment, without handling device external secrets.

The following enrollment phase basically establishes a mutual authenticated TLS connection (using the vendor's device certificate) over which an SCR attribute request may be sent, followed by the CSR generated by the enrolling IED. The CSR itself is enhanced with the TLS session identifier to achieve a binding to the underlying transport. More information about the provisioning using EST can be found in [18].

When deployed in the operational environment, IEDs may not be pre-configured with information about the deployment environment. Hence, an intermediate component is used to enhance the CSR with additional information about the deployment environment before it is forwarded to the RA/CA. This information is not available at, or provided by the sender of the CSR itself.

The following section describes such an enhancement of the CSR communication on the way from the devices to certification server. This enhancement is proposed to provide additional information about the environment in which the device is deployed. Such information can either be contained in the certificate to be issued, or associated with the device certificate by other means, like a central configuration database. This approach helps identifying, e.g., a physical movement of components, or devices to other locations. Hence, key material valid in one location may not be misused in a different location. Moreover, the approach also enhances the options for asset management, by providing fine-grained information already during the authentication processes, employing the enhanced certificate.

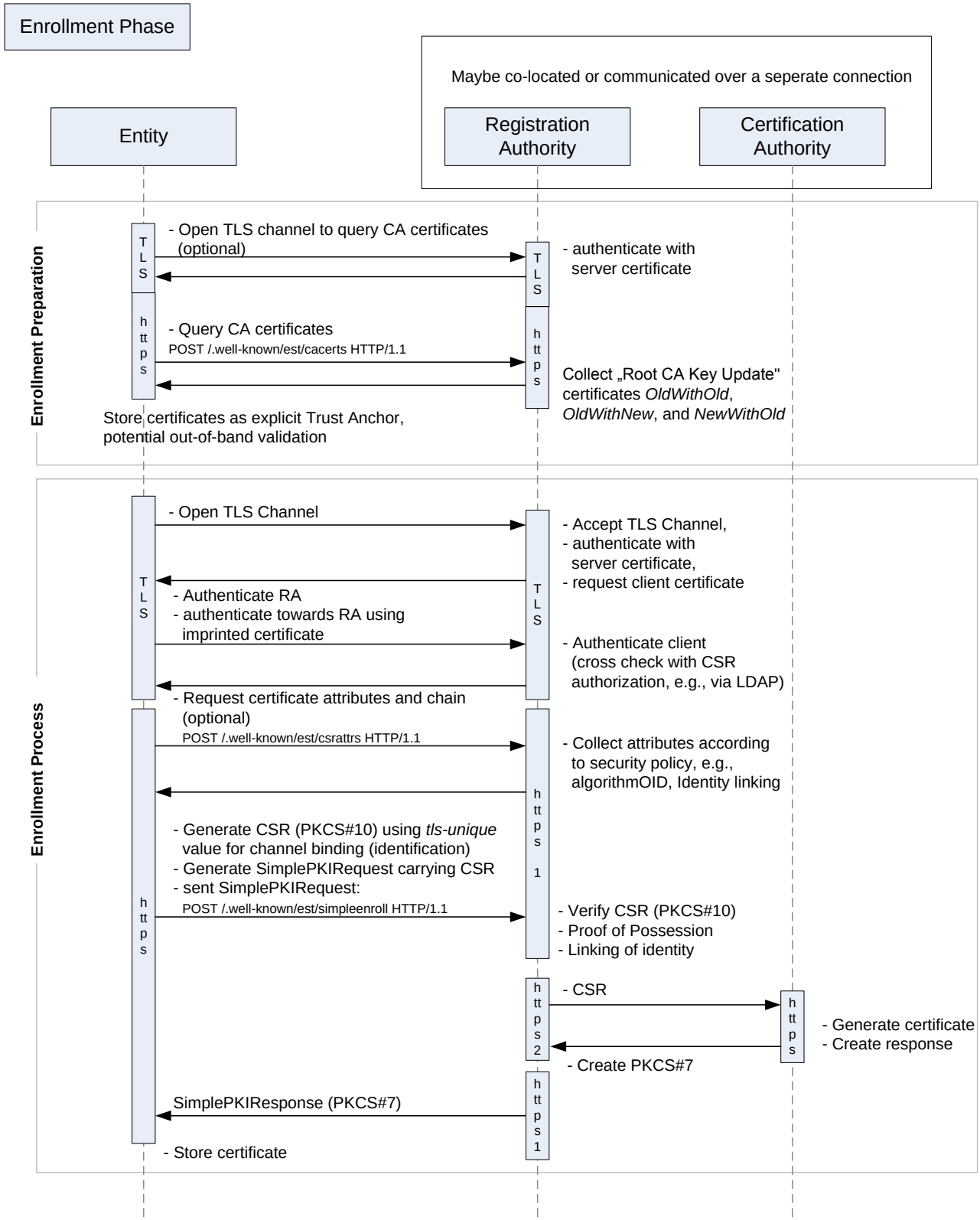


Figure 11. Communication flow hen using Enrollment over Secure Transport [16]

VII. Enhancing Certificate Enrollment with Purpose Binding

This section outlines the introduction of an additional network component to extend a CSR with additional information. Such additional information is encoded as additional attribute added to the original CSR as sent by the device. This attribute indicates the context, or other deployment specific information, to be added to either the certificate, or the configuration database. The original CSR from the device is left untouched including the signature, to preserve the proof of possession of the private key. Note that the original CSR already carries information to enable the issuing RA/CA to identify the device. This can be achieved by using manufacturer installed key material, which can either be used to wrap the CSR providing a digital signature or to establish a connection with the issuer, which invokes this key material. In both cases, the relying party needs to possess the device related certificate material in advance, to be able to perform the authentication.

The enhancement of the original CSR is achieved by adding a Certificate Attribute Intermediary (CAI) along the CSR communication path. The CAI adds at least one attribute to the original CSR (as stated without otherwise manipulating the original CSR). The additional attribute acknowledges additional information about the operating environment. This additional information may be the membership of the CSR sender (device) to a dedicated zone, or group, or to a dedicated location, either on a geographical base, or on an organizational base. Moreover, the CAI may already check the CSR (like an RA), and signal this information also in the attribute. The CAI may add information about intended usage restrictions of the certificate, depending on the device type, and the security policy. This information can be part of the engineering information, which must then be available at the CAI. The CAI may also request that the certificate is issued using a dedicated signature algorithm.

The attribute and the CSR build the Extended Certificate Request (ECR). The ECR is protected by a cryptographic checksum, binding the attributes to the original CSR. Ideally, this is a digital signature of the CAI. This could be realized as PKCS#7 structure [19], or as XML structure, but may also be a symmetric checksum, involving a shared secret between the CAI, and the CA. If the RA, and CA are separate entities, the CAI may be co-located with a local RA. After successful verification, the additional information from the attribute is included in the X.509 certificate within a certificate extension.

Depending on the applied enrollment protocol the ECR may be transmitted via a TLS protected communication path using, e.g., HTTP POST, HTTP GET, or as REST, or SOAP message.

Figure 12 depicts the on path enhancement of a CSR with attributes *aa1, …, aa3*. Also shown are potential functions to be performed by the CAI (e.g., CSR checking), and the enhanced functions on the RA/CA side.

In a substation automation environment, the CAI can be part the substation controller, or the remote access server as the central ingress, and egress point of the substation. This is depicted in Figure 13. The different steps describe the single steps for the ECR processing. Note that the prerequisite is the availability of the central RA/CA root certificate in the IED.

The following steps, which are shown in Figure 13 are performed for the initial enrollment of an energy automation device IED:

1. Generation of key material (public/private key), generation of the CSR within the IED
2. Send local generated CSR to Remote Access Server
3. Verification of CSR through Remote Access Server. Remote Access Server acts as CAI. Generation of attributes, and ECR. Send ECR to central RA/CA server of the distribution network operator.
4. Verification of ECR signature through central RA/CA, verification of attributes (installation information, etc.); optional verification of original CSR
5. In case of successful verification device specific certificate will be generated, and send to the remote access server of the substation.
6. Forwarding of certificate to IED
7. Local automated installation of certificate upon receiving, and successful signature verification against local RA/CA root certificate.

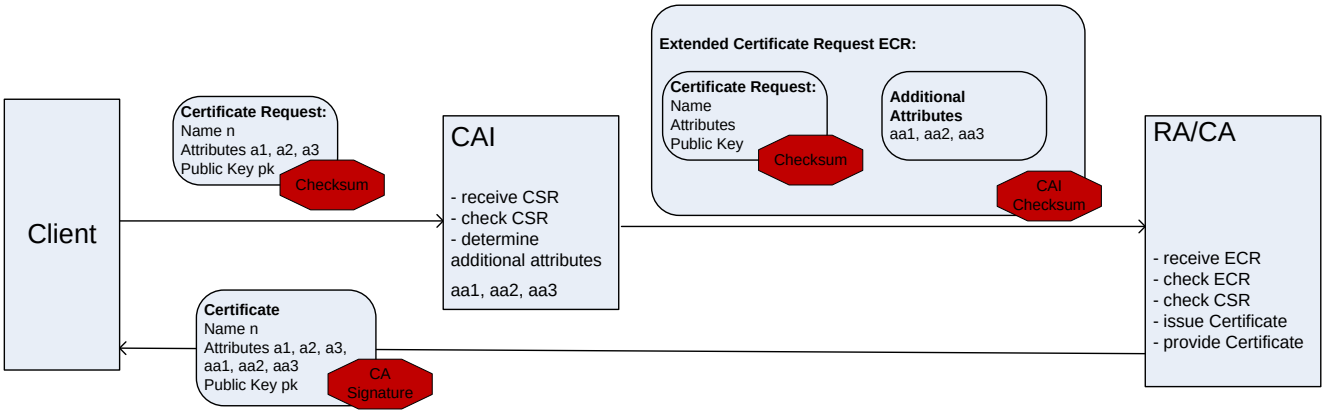


Figure 12. Realization option for on path CSR enhancement with attributes characterizing the deployment environment

There are certain prerequisites to have an automated certificate enrollment. First of all the devices should be able to generate a key pair and also to generate the CSR. This already involves cryptographic computations, which the devices need to be capable of. From an infrastructure perspective, the device identity needs to be known upfront. This can be done very efficiently, if the device already posses a certificate and corresponding private key. The certificate (or a serial number and issuer or a certificate fingerprint) needs to be known by the issuing CA and potentially by the CAI. This key material can then be used in the CSR communication process to authenticate the applying device. The device can authenticate itself by providing a digital signature, either for the CSR wrapper or in the context of the utilized communication protocol, e.g., EST.

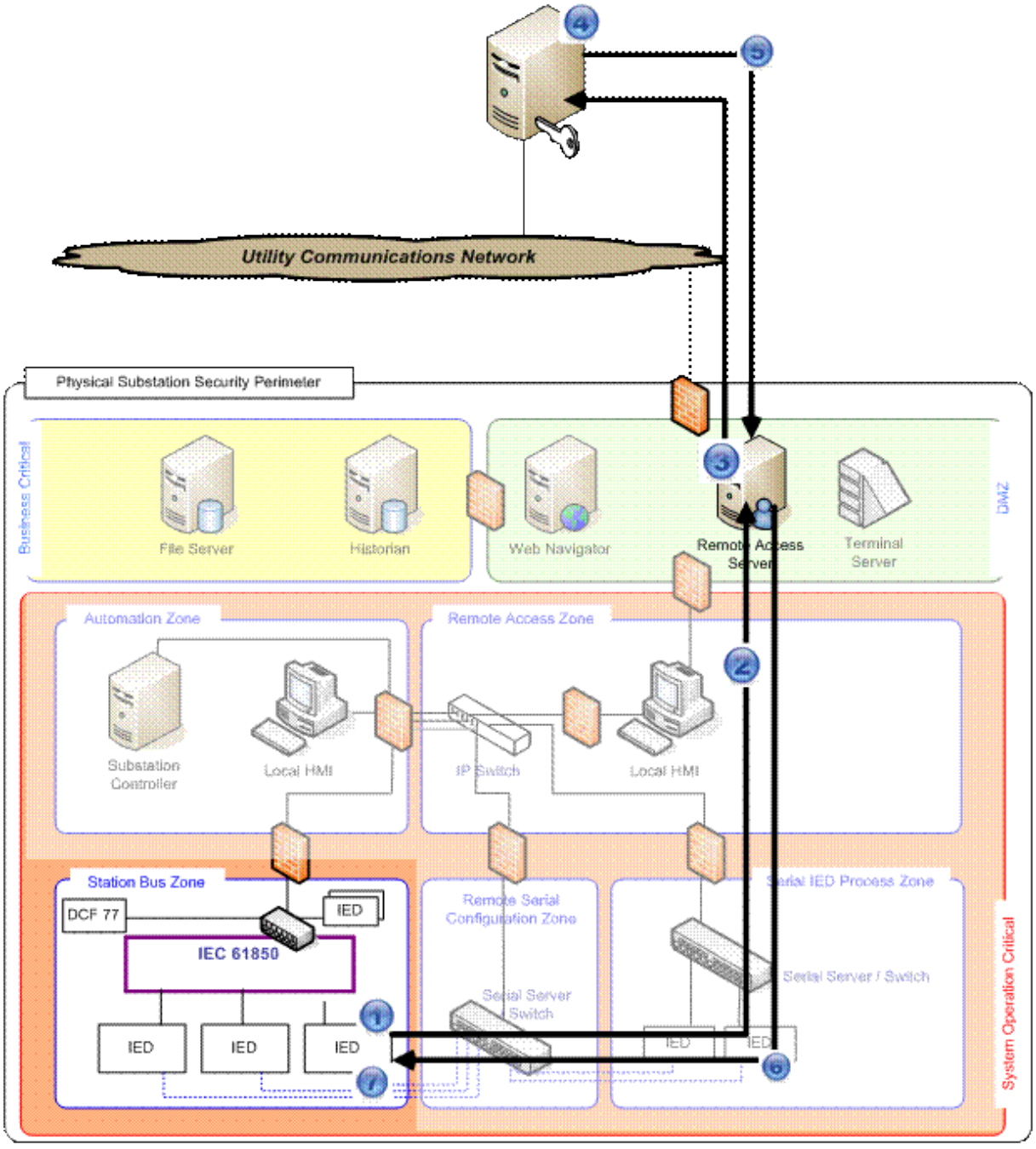


Figure 13. Enhancement of the CSR path in a substation

In case of EST as enrollment protocol, it would be necessary that the remote access server acts as local registration authority (LRA) to enable the enhancement of the device CSR. As EST defines the binding of the CSR to the TLS session, by including the TLS session identifier (*tls_unque*) into the CSR, an intermediate component must be identifiable as trusted. This is enabled by a specific key usage extension in the LRA's certificate called *id-kp-cmcRA*.

Note that this paper concentrates on the concept of the CSR enhancement. Implementations are not finished yet.

VIII. Conclusions and Outlook

This paper described security enhancements for energy automation systems involved in standard substation communication but also in smart grid / smart energy scenarios like the integration of decentralized energy resources (DER). The cryptographic protection of control communication requires that cryptographic keys, and certificates are provisioned on energy automation field-level devices. Manual configuration would not scale to the huge number of devices, and be prone to configuration errors. Therefore, automatic configuration of automation devices is required not only during the operation, and service, but especially for the initial device enrollment. To ensure the correct configuration of cryptographic device credentials, information is required at which location a specific device has been installed. An additional network element has been described that trustfully enhances a certificate signing request issued by an automation device with information on the network segment in which the device has been installed. This allows the CA to issue a device certificate that is bound to the operational zone of the device ("location"). Moreover, additional information for the CSR processing can also be provided. A relying device towards which the considered device authenticates using this zone-bound certificate, can verify whether the device belongs to the own zone. This ensures that an automatically provisioned device is operable using the established configuration only within the corresponding zone. When the device is relocated, or put out of service, its device certificate cannot be misused, e.g., in other zones.

Standardization is currently ongoing in the context of ISO/IEC62351-9, which defines interoperable means for automatic device credential management for energy automation equipment. The new approach described in this paper enhances the current credential management approach, and will be proposed for to be considered in future energy automation security standards. While applicable in the context of energy automation standards, the CSR enhancement may also be standardized in the context of the actual enrollment standards, as pointed out in Section VI.

References

[1] R. Falk and S.Fries, "Purpose-bound Certificates in Automation Environments," Proc. IARIA Internet 2014, June 2014, ISBN 78-1-61208-349-0, pp. 29-33, www.thinkmind.org/index.php?view=article&articleid=internet_2014_2_10_40022), [retrieved April 2015]

[2] S. Fries and R. Falk, "Efficient Multicast Authentication in Energy Environments," Proc. IARIA Energy 2013, March 2013, ISBN 978-1-61208-259-2, pp. 65-71, http://www.thinkmind.org/download.php?articleid=energy_2013_3_30_40056 [retrieved Dec. 2014].

[3] M. Felser, "Real-time Ethernet – industry prospective," Proc. IEEE, vol. 93, no.6, June 2005, pp. 1118-1128, http://www.felser.ch/download/FE-TR-0507.pdf [retrieved: Dec. 2014] .

[4] IEC 61850-5 – "Communication requirements for functions and device models," July 2003, http://www.iec.ch/smartgrid/standards/ [retrieved: Jan. 2015].

[5] "Efficient Energy Automation with the IEC 61850 Standard Application Examples," Siemens AG, December 2010, http://www.energy.siemens.com/mx/pool/hq/energy-topics/standards/iec-61850/Application_examples_en.pdf [retrieved: Dec. 2014].

[6] IEC 62351-x Power systems management and associated information exchange – Data and communication security, http://www.iec.ch/smartgrid/standards/[retrieved: Jan. 2014].

[7] P. Saint-Andre, "Extensible Messaging and Presence Protocol (XMPP): Core," RFC 6120, https://tools.ietf.org/html/rfc6120 [retrieved: Jan. 2014].

[8] T. Dierks and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2," RFC 5246, Aug. 2008, http://tools.ietf.org/html/rfc5246 [retrieved: Jan. 2015].

[9] B. Weiss, S. Rowles, and T. Hardjono, "The Group Domain of Interpretation," RFC 6407, Oct. 2011, http://tools.ietf.org/html/rfc6407 [retrieved: Jan. 2015].

[10] D. Cooper et al., "Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile," RFC 5280, May 2008, http://tools.ietf.org/html/rfc5280 [retrieved: Jan. 2015].

[11] A. Lakshminarayanam and J. Zhou, "FlexiCert: X.509 Identity and Attribute Certificates," Proceedings of DEXA 2003, ISBN 1529-4188/03 [retrieved: Jan. 2015].

[12] M. Nystrom and B. Kaliski, "PKCS #10: Certification Request Syntax Specification," RFC 2986, Nov. 2000, http://tools.ietf.org/html/rfc2986 [retrieved: Jan. 2015].

[13] M. Pritikin, A. Nourse, and J. Vilhuber, "Simple Certificate Enrolment Protocol," Internet Draft, Sep. 2011, http://tools.ietf.org/html/draft-nourse-scep-23 [retrieved: Jan. 2015].

[14] J.Schaad and M.Myers, "Certificate Management over CMS," RFC 5272, June 2008, http://tools.ietf.org/search/rfc5272 [retrieved: Jan. 2015].

[15] C. Adams, S. Farrell, T. Kause, and T. Mononen, "Internet X.509 Public Key Infrastructure Certificate Management Protocol (CMP)," RFC 4210, Sep. 2005, http://tools.ietf.org/html/rfc4210 [retrieved: Jan. 2015].

[16] M. Pritikin, P. Yee, and D. Harkins, "Enrollment over Secure Transport," RFC 7030, Oct. 2013, http://tools.ietf.org/html/rfc7030 [retrieved: Jan. 2015].

[17] XML Key Management Specification, http://www.w3.org/TR/xkms2/

[18] J. Foley, "Provisioning X.509 Certificates Using RFC 7030," Linux Journal, September 2014, http://www.linuxjournaldigital.com/linuxjournal/september_2014/?pg=62&pm=2&u1=friend#pg62 [retrieved: Dec. 2014].

[19] B. Kaliski, "PKCS#7 Cryptographic Message Syntax Version 1.5," RFC2315, March 1998, http://tools.ietf.org/html/rfc2315 [retrieved: Dec. 2014].

A Benchmark Survey of Rigid 3D Point Cloud Registration Algorithms

Ben Bellekens*, Vincent Spruyt*, Rafael Berkvens*, Rudi Penne†, and Maarten Weyn*
*CoSys-Lab, Faculty of Applied Engineering
University of Antwerp, Belgium
Email: {ben.bellekens, rafael.berkvens, maarten.weyn}@uantwerpen.be and v.spruyt@ieee.org
†Op3Mech, Faculty of Applied Engineering, Dept. of Mathematics
University of Antwerp, Belgium
Email: rudi.penne@uantwerpen.be

***Abstract*—Advanced user interface sensors are able to observe the environment in three dimensions with the use of specific optical techniques such as time-of-flight, structured light or stereo vision. Due to the success of modern sensors, which are able to fuse depth and color information of the environment, a new focus on different domains appears. This survey studies different state-of-the-art registration algorithms, which are able to determine the motion between two corresponding 3D point clouds. This survey starts from a mathematical field of view by explaining two deterministic methods, namely Principle Component Analysis (PCA) and Singular Value Decomposition (SVD), towards more iteratively methods such as Iterative Closest Point (ICP) and its variants. We compare the performance of the different algorithms to their precision and robustness based on a real world dataset. The main contribution of this survey consists of the performance benchmark that is based on a real world dataset, which includes 3D point clouds of a Microsoft Kinect camera, and a mathematical overview of different registration methods, which are commonly used in applications such as simultaneous localization and mapping, and 3D-scanning. The outcome of our benchmark concludes that the ICP point-to-surface method is the most precise algorithm. Beside the precision, the result for the robustness we can conclude that a combination of applying a ICP point-to-point method after an SVD method gives the minimum error.**

Keywords–3D point cloud; PCL; Kinect camera; 3D Fine registration; rigid transformation; survey paper; Robustness; Precision; SLAM

I. INTRODUCTION

This article, which is an extended version of the conference paper [1], contains new results that defines the robustness and the precision of the different registration algorithms.

With the advent of inexpensive depth sensing devices, robotics, computer vision and ambient application technology research has shifted from 2D imaging and Laser Imaging Detection And Ranging (LIDAR) scanning towards real-time reconstruction of the environment based on 3D point cloud data. On the one hand, there are structured light based sensors such as the Microsoft Kinect and Asus Xtion sensor, which generate a structured point cloud, sampled on a regular grid, and on the other hand, there are many time-of-flight based sensors such as the Softkinetic Depthsense camera, which yield an unstructured point cloud. These point clouds can either be used directly to detect and recognize objects in the environment where ambient technology is been used, or can be integrated over time to completely reconstruct a 3D map of the camera's surroundings [2], [3], [4]. However, in the latter case, point clouds obtained at different time instances need to be aligned, a process that is often referred to as registration. Registration algorithms are able to estimate the ego-motion of a robot by calculating the transformation that optimally maps two point clouds, each of which is subject to camera noise.

Registration algorithms can be classified coarsely into rigid and non-rigid approaches. Rigid approaches assume a fixed rigid environment such that a homogeneous transformation can be modelled using only 6 Degrees Of Freedom (DOF). On the other hand, non-rigid methods are able to cope with articulated objects or soft bodies that change shape over time. Additionally, registration algorithms can be classified into coarse and fine approaches. Coarse registration approaches compute an initial geometric alignment whereas fine registration approaches compute a transformation that can register two point clouds precisely. A combination of coarse and fine registration algorithms is often used in applications to reduce the number of iterations while an optimal alignment still occurs.

Registration algorithms are used in different fields and applications, such as 3D object scanning, 3D mapping, 3D localization and ego-motion estimation or human body detection. Most of these state-of-the-art applications employ either a simple Singular Value Decomposition (SVD) [5] or Principal Component Analysis (PCA) based registration, or use a more advanced iterative scheme based on the Iterative Closest Point (ICP) algorithm [6]. Recently, many variants on the original ICP approach have been proposed, the most important of which are non-linear ICP [7], and generalized ICP [8]. These are explained and discussed in this publication.

To our knowledge, a general discussion of each of the above methods that are applied in a real world scenario where environment data is been acquired with a 3D sensor is not available in literature. Salvi *et al.* presented a survey article, which gives an overall view of coarse and fine registration methods that are able to register range based images [9]. But they presented a performance comparison based on synthetic data and real data that was recorded by a laser scanner.

The choice of an algorithm generally depends on several important characteristics such as accuracy, computational complexity, and convergence rate, each of which depends on the application of interest. Moreover, the characteristics of most registration algorithms heavily depend on the data used, and thus on the environment itself. As a result, it is difficult to compare these algorithms data independently. Therefore, in this paper we discuss the mathematical foundations that are common to the most widely used 3D registration algorithms, and we compare their robustness and precision in a real world situations.

This paper is outlined as follows: Section II briefly discusses several important application domains of 3D registration algorithms. In Section III, rigid registration is formulated as a least square optimization problem. Section IV explains the most important rigid registrations algorithms, which are PCA, SVD, ICP point-to-point, ICP point-to-surface, ICP non-linear and Generalized ICP. Finally, Section V provides a discussion of the precision and the robustness of each of these methods in a real world setting. Section VI concludes the paper.

II. Application Domains

Important application domains of both rigid and non-rigid registration methodologies are robotics, healthcare, astrophotography, and more. In these applications, the common goal is to determine the position or pose of an object with respect to a given viewpoint. Whereas rigid transformations are defined by 6 DOF, non-rigid transformations allow a higher number of DOF in order to cope with non-linear or partial stretching or shrinking of the object [10]. Following subsections will give an overview of the robotic applications and healthcare applications where 3D rigid registration methods are being applied.

A. Robotics

Since the introduction of inexpensive depth sensors such as the Microsoft Kinect camera, great progress has been made in the robotic domain towards Simultaneous Localization And Mapping (SLAM) [11], [12], [13], [14]. The reconstructed 3D occupancy grid map is represented by a set of point clouds, which are aligned by means of registration and can be used for techniques such as obstacle avoidance, map exploration and autonomous vehicle control [4], [15], [16]. Furthermore, depth information is often combined with a traditional RGB camera [3], [17] in order to greatly facilitate real-world problems such as object detection in cluttered scenes, object tracking and object recognition [18]. The main goal in robotic applications is to develop a robust, precise and accurate algorithm that can execute almost at real-time. In order to reach this goal much research is nowadays focusing toward graphical processing unit (GPU) and multicore processing, which enables the execution of many computation task during one timeslot on multiple processing cores [19], [20].

B. Healthcare

Typical applications of non-rigid registration algorithms can be found in healthcare, where a soft-body model often needs to be aligned accurately with a set of 3D measurements. Applications are cancer-tissue detections, hole detection, artefact recognition, etc. [10], [21]. Similarly, non-rigid transformations are used to obtain a multi-modal representation of a scene, by combining magnetic resonance imaging (MRI), computer tomography (CT), and positron emission tomography PET volumes into a single 3D model [10].

III. Definitions

In this section, we briefly introduce the least-squares optimization problem and discuss the concept of homogeneous transformations since these form the basis of 3D registration algorithms.

Rigid registration can be approached by defining a cost function that represents the current error, which indicates how well two point clouds overlap. This cost function is then minimized using common optimization techniques. If the distance between corresponding points in each 3D point cloud needs to be minimized, this can be simplified to a linear least-squares minimization problem by representing each point using homogeneous coordinates.

A. Homogeneous transformations

A homogeneous transformation in three dimensions is specified by a 4×4 projective transformation matrix [22]. This matrix is used to project each point in Cartesian space with respect to a specific viewpoint. Since we use (moving) rigid orthonormal reference frames, we can restrict our considerations to rigid transformations. In the following, let $\tilde{\mathbf{v_1}} = (x_1, y_1, z_1, 1)^\intercal$ be standard homogeneous coordinates of a point in an orthonormal base defined by viewpoint one, and let $\tilde{\mathbf{v_2}} = (x_2, y_2, z_2, 1)^\intercal$ be standard homogeneous coordinates of the same point in an orthonormal base defined by viewpoint two. Then it is possible to express $\tilde{\mathbf{v_2}}$ relative to the base of viewpoint one as $T\tilde{\mathbf{v_1}} = \tilde{\mathbf{v_2}}$, where T is a Euclidean transformation matrix defined by (1).

$$T = \begin{pmatrix} r_{1,1} & r_{1,2} & r_{1,3} & t_1 \\ r_{2,1} & r_{2,2} & r_{2,3} & t_2 \\ r_{3,1} & r_{3,2} & r_{3,3} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{1}$$

The transformation matrix shown by (1) consists of a 3×3 rotation matrix (2),

$$R = \begin{pmatrix} r_{1,1} & r_{1,2} & r_{1,3} \\ r_{2,1} & r_{2,2} & r_{2,3} \\ r_{3,1} & r_{3,2} & r_{3,3} \end{pmatrix} \tag{2}$$

and the column vector $\vec{t} = (\vec{t_1}, \vec{t_2}, \vec{t_3})^\intercal$ representing a translation. Because the nine entries of the rotation matrix can be generated by three parameters (e.g., the Euler angles), we conclude that a rigid transformation has six DOF.

B. Least-Squares Minimization

A rigid transformation is defined by only 6 DOF, whereas many noisy observations, i.e., point coordinates, are available. Therefore, the number of parameters of any cost function for this problem is much smaller than the number of equations, resulting in an ill-posed problem that does not have an exact solution. A well known technique to obtain an acceptable solution in such case, is to minimize the square of the residual error. This approach is called least-squares optimization and is often used for fitting and regression problems.

Whereas a linear least-squares problem can be solved analytically, this is often not the case for non-linear least-squares optimization problems. In this case, an iterative approach can be used by iteratively exploring the search space of all possible solutions in the direction of the gradient vector of the cost function. This is illustrated by Figure 1, where the cost function $f(d)$ of the ICP registration algorithm is minimized iteratively. The cost function in this case represents the sum of the squared Euclidean distances, defined by the rotation and the translation, between all corresponding points of two point cloud viewpoints.

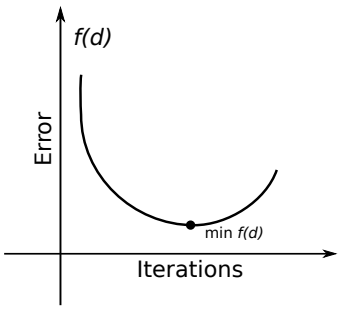


Figure 1. ICP Least square approach.

IV. Registration Algorithms

Both rigid and non-rigid registration algorithms can be further categorized into pairwise registration algorithms and multiview registration methods. Pairwise registration algorithms calculate a rigid transformation between two subsequent point clouds while the multi-view registration process takes multiple point clouds into account to correct for the accumulated drift that is introduced by pairwise registration methods.

In the next sections, we discuss five widely used rigid registration algorithms. Each of these methods tries to estimate the optimal rigid transformation that maps a source point cloud on a target point cloud. Both PCA alignment and SVD are pairwise registration methods based on the covariance matrices and the cross correlation matrix of the point clouds, while the ICP algorithm and its variants are based on iteratively minimizing a cost function that is based on an estimate of point correspondences between the point clouds. The selected correspondences will determine the quality of how the final transformation fits the source point cloud to the target point cloud.

A. Principal Component Analysis

PCA is often used in classification and compression techniques to project data on a new orthonormal basis in the direction of the largest variance [23]. The direction of the largest variance corresponds to the largest eigenvector of the covariance matrix of the data, whereas the magnitude of this variance is defined by the corresponding eigenvalue.

Therefore, if the covariance matrix of two point clouds differs from the identity matrix, a rough registration can be obtained by simply aligning the eigenvectors of their covariance matrices. This alignment is obtained as follows.

First, the two point clouds are centered such that the origins of their original bases coincide. Point cloud centering simply corresponds to subtracting the centroid coordinates from each of the point coordinates. The centroid of the point cloud corresponds to the average coordinate and is thus obtained by dividing the sum of all point-coordinates by the number of points in the point cloud.

Since registration based on PCA simply aligns the directions in which the point clouds vary the most, the second step consists of calculating the covariance matrix of each point cloud. The covariance matrix is an orthogonal 3×3 matrix, the diagonal values of which represent the variances while the off-diagonal values represent the covariances.

Third, the eigenvectors of both covariance matrices are calculated. The largest eigenvector is a vector in the direction of the largest variance of the 3D point cloud and, therefore, it represents the point cloud's orientation. In the following, let A be the covariance matrix, let $\vec{v}$ be an eigenvector of this matrix, and let λ be the corresponding eigenvalue. The eigenvalues decomposition problem is then defined as:

$$A\tilde{\mathbf{x}} = \lambda\tilde{\mathbf{x}} \tag{3}$$

and further reduces to:

$$\tilde{\mathbf{x}}(A - \lambda I) = 0. \tag{4}$$

It is clear that (4) only has a non-zero solution if $A - \lambda I$ is singular and, consequently, if its determinant equals zero:

$$\det(A - \lambda I) = 0 \tag{5}$$

The eigenvalues can simply be obtained by solving (5), whereas the corresponding eigenvectors are obtained by substituting the eigenvalues into (3).

Once the eigenvectors are known for each point cloud, registration is achieved by aligning these vectors. In the following, let matrix T_t^y represent the transformation that would align the largest eigenvector t of the target point cloud with the y-axis. Let matrix T_y^s represent the transformation that would align the largest eigenvector s of the source point cloud with the y-axis. Then the final transformation matrix T_t^s that aligns the source point cloud with the target point cloud can be obtained easily, as illustrated by Figure 2.

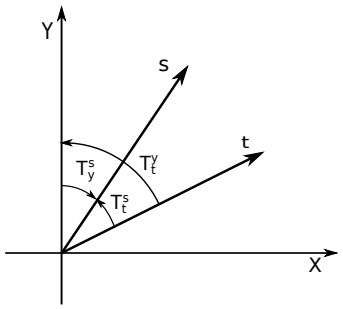


Figure 2. PCA alignment from source to target.

Finally, the centroid of the target data is added to each of the transformed coordinates to translate the aligned point cloud, such that its center corresponds to the center of the target point cloud.

B. Singular Value Decomposition

PCA based registration simply aligns the directions of the largest variance of each point cloud and, therefore, it does not minimize the Euclidean distance between corresponding points of the datasets. Consequently, this approach is very sensitive to outliers and only works well if each point cloud is approximately normally distributed.

However, if point correspondences between the two point clouds are available, a more robust approach would be to directly minimize the sum of the Euclidean distances between these points. This corresponds to a linear least-squares problem that can be solved robustly using the SVD method [5].

Based on the point correspondences, the cross correlation matrix M between the two centered point clouds can be calculated, after which the eigenvalue decomposition is obtained as follows:

$$M = USV^{\intercal} \tag{6}$$

The optimal solution to the least-squares problem is then defined by rotation matrix R as:

$$R_t^s = UV^{\intercal} \tag{7}$$

and the translation from target point cloud to source point cloud is defined by:

$$\tilde{\mathbf{t}} = \tilde{\mathbf{c}_\mathbf{s}} - R_t^s \tilde{\mathbf{c}_\mathbf{t}} \tag{8}$$

C. Iterative Closest Point

Whereas the SVD algorithm directly solves the least-squares problem, thereby assuming perfect data, Besl and Mc. Kay [6] introduced a method that iteratively disregards outliers in order to improve upon the previous estimate of the rotation and translation parameters. Their method is called 'ICP' and is illustrated conceptually in Figure 3.

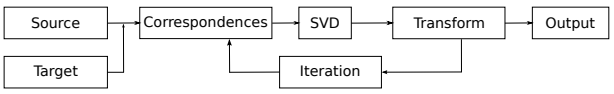


Figure 3. ICP overview scheme.

The input of the ICP algorithm consists of a source point cloud and a target point cloud. Point correspondences between these point clouds are defined based on a nearest neighbour approach or a more elaborate scheme using geometrical features or color information. SVD, as explained in the previous section, is used to obtain an initial estimate of the affine transformation matrix that aligns both point clouds. After transformation, this whole process is repeated by removing outliers and redefining the point correspondences.

Two widely used ICP variants are the ICP point-to-point and the ICP point-to-surface algorithms. These approaches only differ in their definition of point correspondences and are described in more detail in the next sections.

1) ICP point-to-point: The ICP point-to-point algorithm was originally described in [2] and simply obtains point correspondences by searching for the nearest neighbour target point $\vec{q_i}$ of a point $\vec{p_j}$ in the source point cloud. The nearest neighbour matching is defined in terms of the Euclidean distance metric:

$$\hat{i} = \arg\min_i \|\vec{p_i} - \vec{q_j}\|^2, \tag{9}$$

where $i \in [0, 1, ..., N]$, and N represents the number of points in the target point cloud.

Similar to the SVD approach discussed in Section IV-B, the rotation R and translation $\vec{t}$ parameters are estimated by minimizing the squared distance between these corresponding pairs:

$$\hat{R}, \hat{\vec{t}} = \arg\min_{R,\vec{t}} \sum_{i=1}^{N} \|(R\vec{p_i} + \vec{t}) - \vec{q_i}\|^2 \tag{10}$$

ICP then iteratively solves (9) and (10) to improve upon the estimates of the previous iterations. This is illustrated by Figure 4, where surface s is aligned to surface t after n ICP iterations.

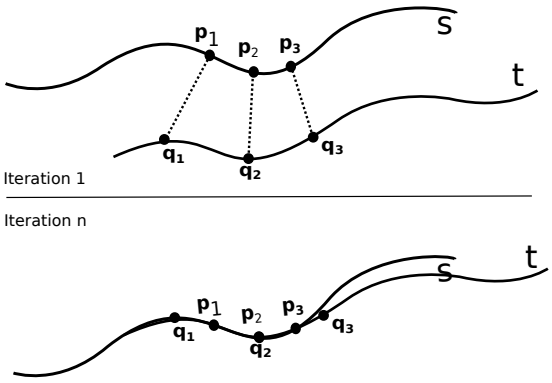


Figure 4. ICP alignment based on a point to point approach.

2) ICP point-to-surface: Due to the simplistic definition of point correspondences, the ICP point-to-point algorithm proposed by [24] is rather sensitive to outliers. Instead of directly finding the nearest neighbour to a source point $\vec{p_j}$ in the target point cloud, one could take the local neighbourhood of a correspondence candidate $\vec{q_i}$ into account to reduce the algorithm's sensitivity to noise.

The ICP point-to-surface algorithm assumes that the local neighbourhood of a point in a point cloud is co-planar. This local surface can then be defined by its normal vector $\vec{n}$, which is obtained as the smallest eigenvector of the covariance matrix of the points that surround correspondence candidate $\vec{q_i}$.

Instead of directly minimizing the Euclidean distance between corresponding points, we can then minimize the scalar projection of this distance onto the planar surface defined by the normal vector $\vec{n}$:

$$\hat{R}, \hat{\vec{t}} = \arg\min_{\hat{R},\hat{\vec{t}}} \left(\sum_{i=1}^{N} \|((R\vec{p_i} + \vec{t}) - \vec{q_i})\vec{n_i}\| \right) \tag{11}$$

This is illustrated more clearly in Figure 5.

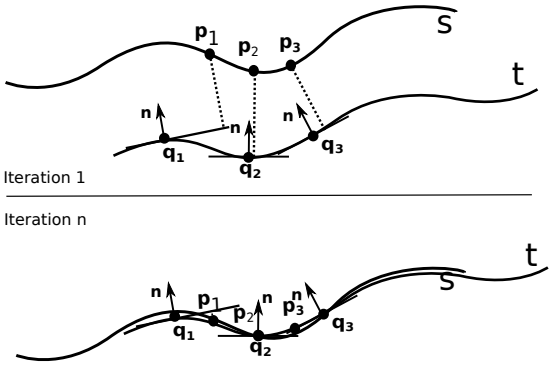


Figure 5. ICP alignment based on a point to surface approach.

3) ICP non-linear: Both the point-to-point and point-to-surface ICP approaches defined a differentiable, convex, squared cost function, resulting in a simple linear least-squares optimization problem, known as a L2-optimization, that can be solved numerically using SVD. However, L2-optimization is known to be highly sensitive to outliers because the residuals are squared. An approach that solves this problem is known as L1-optimization, where the sum of the absolute value of the residuals is minimized instead of the square. However, the L1 cost function is non-differentiable at the origin, which makes it difficult to obtain the optimal solution.

As a compromise between L1 and L2 optimization, the so called Huber loss function can be used as shown by (12). The Huber loss function is quadratic for small values and thus behaves like an L2 problem in these cases. For large values, however, the loss function becomes linear and, therefore, it behaves like an L1 cost function. As Figure 6 shows differentiation between the Huber-Loss function by the green curve, the blue curve shows the L2 quadratic function. Moreover, the Huber loss function is smooth and differentiable, allowing traditional numerical optimization methods to be used to efficiently traverse the search space.

$$e(n) = \begin{cases} n^2/2 & \text{if } |n| \leq k \\ k|n| - k^2/2 & \text{if } |n| > k \end{cases} \quad (12)$$

where k is an empirically defined threshold and n is the distance measure.

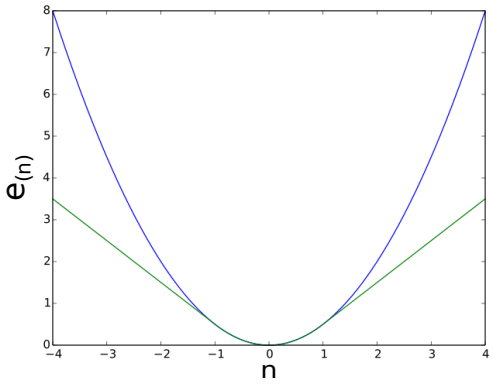


Figure 6. Huber Loss function.

The ICP non-linear algorithm uses the Huber loss function instead of a naive squared loss function to reduce the influence of outliers:

$$\hat{R}, \vec{\hat{t}} = \arg\min_{\hat{R}, \vec{t}} \sum_{i=1}^{N} e^2(n) \quad (13)$$

where

$$n = \|(R\vec{p} - \vec{t}) - \vec{q}\| \quad (14)$$

To obtain the optimal estimates $\hat{R}, \vec{\hat{t}}$ in (13), the Levenberg-Marquardt algorithm (LMA) [7] is used. The LMA method is an iterative procedure similar to the well known gradient descent and Gauss-Newton algorithms, which can quickly find a local minimum in non-linear functions.

4) Generalized ICP: A major disadvantage of the traditional point-to-point ICP algorithm, is that it assumes that the source point cloud is taken from a known geometric surface instead of being obtained through noisy measurements. However, due to discretization errors it is usually impossible to obtain a perfect point-to-point matching even after full convergence of the algorithm. The point-to-surface ICP algorithm relaxes this constraint by allowing point offsets along the surface, in order to cope with discretization differences. However, this approach still assumes that the source point cloud represents a discretized sample set of a known geometric surface model since offsets along the surface are only allowed in the target point cloud.

To solve this, Segal *et al.* [8] proposed the Generalized ICP (GICP) algorithm that performs plane-to-plane matching. They introduced a probabilistic interpretation of the minimization process such that structural information from both the source point cloud and the target point cloud can be incorporated easily in the optimization algorithm. Moreover, they showed that the traditional point-to-point and point-to-surface ICP algorithms are merely special cases of the Generalized ICP framework.

Instead of assuming that the source point cloud is obtained from a known geometric surface, Segal *et al.* assume that both the source point cloud $A = \{\vec{a}_i\}$ and the target point cloud $B = \{\vec{b}_i\}$ consist of random samples from an underlying unknown point cloud $\hat{A} = \{\hat{\vec{a}}_i\}$ and $\hat{B} = \{\hat{\vec{b}}_i\}$. For the underlying and unknown point clouds $\hat{A}$ and $\hat{B}$, perfect correspondences exist, whereas this is not the case for the observed point clouds A and B, since each point $\vec{a}_i$ and $\vec{b}_i$ is assumed to be sampled from a normal distribution such that $\vec{a}_i \sim \mathcal{N}(\hat{\vec{a}}_i, C_i^A)$ and $\vec{b}_i \sim \mathcal{N}(\hat{\vec{b}}_i, C_i^B)$. The covariance matrices C_i^A and C_i^B are unknown. If both point clouds would consist of deterministic samples from known geometric models, then both covariance matrices would be zero such that then $A = \hat{A}$ and $B = \hat{B}$.

In the following, let T be the affine transformation matrix that defines the mapping from $\hat{A}$ to $\hat{B}$ such that $\hat{\vec{b}}_i = T\hat{\vec{a}}_i$. If T would be known, we could apply this transformation to the observed source point cloud A, and define the error to be minimized as $d_i^T = \vec{b}_i - T\vec{a}_i$. Because both $\vec{a}_i$ and $\vec{b}_i$ are assumed to be drawn from independent normal distributions d_i^T, which is a linear combination of $\vec{a}_i$ and $\vec{b}_i$, is also drawn from a normal distribution:

$$d_i^T \sim \mathcal{N}(\hat{\vec{b}}_i - T\hat{\vec{a}}_i, C_i^B + TC_i^A T^\intercal) \quad (15)$$
$$= \mathcal{N}(0, C_i^B + TC_i^A T^\intercal) \quad (16)$$

The optimal transformation matrix $\hat{T}$ is then the transformation that minimizes the negative log-likelihood of the observed errors d_i:

$$\hat{T} = \arg\min_T \sum_i \log\left(p(d_i^T)\right)$$
$$= \arg\min_T \sum_i d_i^{T\intercal} (C_i^B + TC_i^A T^\intercal)^{-1} d_i^T \quad (17)$$

Segal *et al.* showed that both point-to-point and point-to-plane ICP are specific cases of (17), only differing in their choice of covariance matrices C_i^A and C_i^B; If the source point

cloud is assumed to be obtained from a known geometric surface, $C_i^A = 0$. Furthermore, if points in the target point cloud are allowed three degrees of freedom, then $C_i^B = I$. In this case, (18) reduces to:

$$\hat{T} = \arg\min_T \sum_i d_i^{T\intercal} d_i^T = \arg\min_T \sum_i \|d_i^T\|^2, \quad (18)$$

which indeed is exactly the optimization problem that is solved by the traditional point-to-point ICP algorithm. Similarly, C_i^A and C_i^B can be chosen such that obtaining the maximum likelihood estimator corresponds to minimizing the point-to-plane or the plane-to-plane distances between both point clouds.

V. Results & Discussion

In this section, we illustrate the performance of the different registration methods that are based on an iteratively approach. In order to illustrate the performance we tested the precision and the robustness of the different methods. The robustness factor of an algorithm will explain how well an algorithm performs during a period of time on different input parameters. Besides the robustness, the precision factor will clarify how well an algorithm performs on the same input parameter. The results for precision and robustness are based on a set of 3D point clouds that are included in a dataset. All results are generated using the Robot Operating System (ROS) and the Point cloud Library (PCL) [25], [26]. Furthermore, the execution processes of the different methods are calculated by an Asus Zenbook UX32VD, core i7-3517U in combination with 10 GB of RAM-memory.

A. Dataset

The dataset that we used to benchmark the performance is built by a Pioneer-3dx robot and consists of a laser scanner, odometry hardware and 3D point cloud data. The Pioneer-3dx robot is a commonly used robot for academic and research purposes. See Figure 7 for the robot used to build this dataset. To ensure that all sensor measurements have a time-stamp and transformation with respect to the center of the robot, we have used the ROS.

On one hand, ROS is used as a tool to record all sensor measurement including the timestamps, while on the other hand, ROS is used as a platform to schedule the different 3D point clouds based on their timestamps. To reduce the size of the dataset, we decreased the number of point clouds per second. Figure 8 visualizes the dataset by means of an occupancy grid map and a travelled path.

The occupancy grid map is the result of a Rao-Blackwellized particle filter SLAM algorithm with a Bayesian probability distribution [27]. The implementation that we used utilizes the laser range scanner and odometry data to generate an occupancy grid map. However, the location updates are performed by the algorithm are not used to recalculate the travelled path, resulting in a periodically erratic trajectory. To obtain a smooth trajectory, we used the occupancy grid map calculated by the SLAM algorithm to perform adaptive Monte Carlo localization [28]. Because we knew the initial position of the robot, the algorithm did not have to perform global localization, but simply had to track the robot during the complete run. This ensures that location corrections are applied incrementally, resulting in the smooth trajectory. Thus, after the SLAM method has calculated an occupancy grid map, the trajectory was calculated by an Adaptive Monte Carlo Localization approach.

Figure 7. The mobile Pioneer-3dx robot with a mounted Microsoft Kinect Camera, Laser scanner and Sonar sensor.

Figure 8. Occupancy grid map from SLAM approach and the smoothed travelled path

B. Robustness

To measure the robustness of the rigid 3D point cloud registration algorithms, we applied them at various times on different corresponding point clouds and recorded their error and computation time. By averaging over these data points, we obtain information about the robustness of a specific algorithm. We want to compute the robustness of the different rigid registration algorithms so that we can analyze, which algorithm performs best in a real world scenario. We focused on a scenario of mapping an indoor environment to generate a 3D model in which all spatial objects are visible and correctly aligned. We could iterate over all point clouds in our database thanks to the timestamps and playback mechanisms in ROS. Figure 9 shows us a one dimensional axis with vertical marker. Each of these markers represent a 3D point cloud, which was taken at a certain time with respect to the start pose or the beginning of the dataset.

For each set of two point clouds, the fitness score, the averaged and normalized error after registration and alignment between the two point clouds, and the computation time of each algorithm is computed to measure the robustness of the

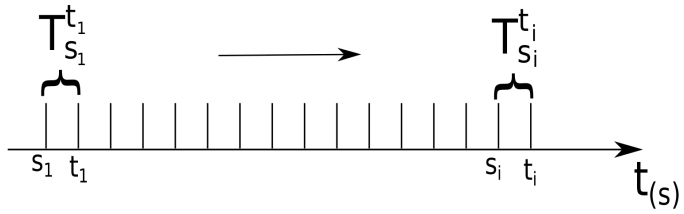


Figure 9. The benchmark robustness scheme includes a set of two 3D point clouds. Each set contains a source point cloud S_i, a target point cloud t_i, and a transformation. Every point cloud is indicated as an individual marker on the time line t.

different algorithms. In this case, there were 165 sets of point cloud pairs or 330 single point clouds.

Figure 10 compares the number of iterations to the averaged fitness score after geometric alignment for each iterative registration process. The result of this correlation can be seen on the green curves, which all converge towards a minimum at 40 iterations. Within this dataset the average of the ICP point-to-point algorithm reaches the lowest minimum in comparison to the other ICP variants. As already stated in the introduction an ICP approach is often used after a coarse registration that can lead to lower minimum. As can be seen in Figure 10, the lowest error value at 40 iterations is SVD_ ICP. This means that a coarse SVD registration has been applied onto the point cloud pair after which an ICP point-to-point is applied. Secondly, the figure shows the computation time for each algorithm at a specific number of iterations. GICP has the worst computation time while ICP point-to-point has the fastest computation time. The reason why ICP point-to-surface is slower than ICP point-to-point is mainly due to the surface normal vector computation. This normal vector computation time could be decreased if the number of nearest neighbour points that should be included onto the surface, is lower. This will change the behaviour, so it will gradually perform more like an ICP point-to-point approach.

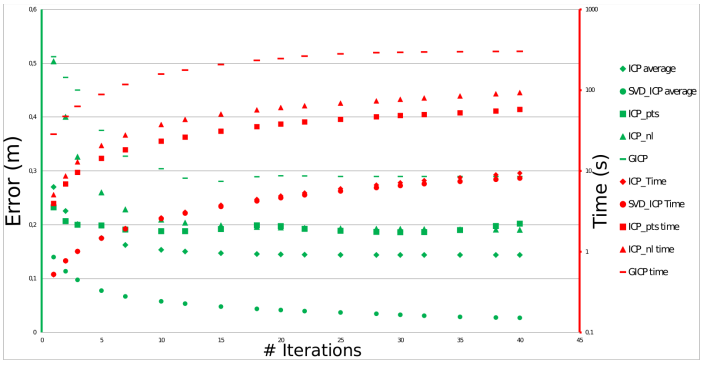


Figure 10. This figure show the comparison between the number of registration iteration and the time logarithmic in red and the comparison between the number of iterations and the fitness score in green for the average of ICP point-to-point (ICP), SVD applied before ICP (SVD_ICP), ICP point-to-surface (ICP_pts), ICP non-linear (ICP_nl) and Generalized ICP (GICP)

The previous paragraph stated the robustness as the average fitness score after alignment while this paraghraph will define the robustness by the sum of the average and the distance of one variance. Thus, the robustness factor is not only the average of each registration method, measured on a set of different 3D point clouds, but it also depending on the variance of the averaged fitness score or how far the fitness score will change over time. This result are visualized in Figure 11. On this graph, the number of iterations is shown on the x-axis and the sum of the average with the distance of one variance onto the y-axis. The robustness of the ICP point-to-surface method is very good due to the constant behavior during the entire dataset. This behavior is normal because the number of new surfaces will not decrease over time whilst two point clouds are being registered. In contrast to the previous method, the robustness of the other ICP approaches will go from worst in the beginning to better at the end due to the many changes in correspondences while registering two point clouds. When applying a coarse registration before an ICP approach the robustness will be much better at convergence than using all other stated methods.

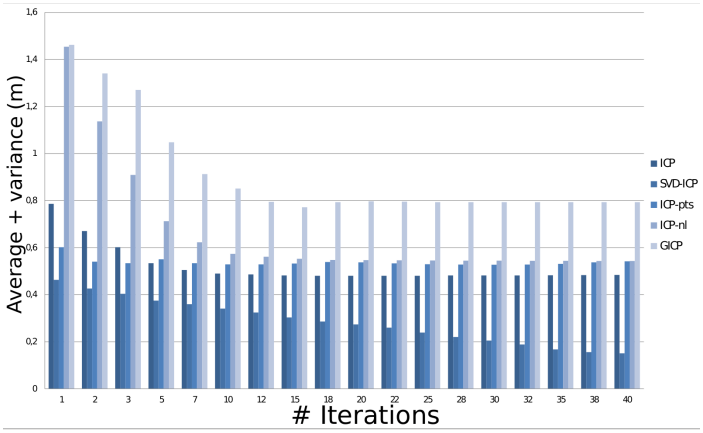


Figure 11. The x-axis represents the number of registration iterations and the y-axis represents the sum of the average and the variance for ICP point-to-point (ICP), SVD applied before ICP (SVD_ICP), ICP point-to-surface (ICP_pts), ICP non-linear (ICP_nl) and Generalized ICP (GICP)

C. Precision

To illustrate the behaviour of the results of the different stated registration algorithms during a certain period of time, the robustness was computed. In order to analyze the precision of the different registration algorithms, the rotation and translation part of the transformation matrix after alignment will be discussed separately. The precision of the different algorithms will illustrate how well they perform on the same two point clouds but with different correspondences. Figure 12 expands the flow that is used to compute the precision of the stated registration algorithms. Depending on the number of precision iterations, more or less subsamples will be computed. Thus, each subsample of the source point cloud will be registered with the target point cloud, which results in a series of alignment transformations that are compliant with the lowest fitness score at 40 iterations. Furthermore, the list of transformations will be divided into a list of rotation matrices and a list of translation matrices. In order to compare the different rotations independently, the 3×3 rotation matrices had to be converted into Euler angles. This means that each rotation around the x, y, and z axes can be represented by yaw, pitch and roll.

The different subsamples of the source point cloud has less points than the initial source point cloud and they are created on set of random indices, which are based on the indices of the source point cloud.

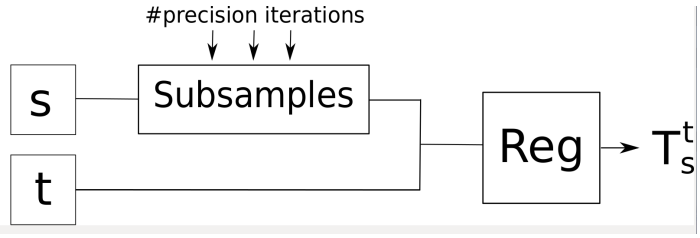


Figure 12. The benchmark precision scheme

To illustrate the precision of the translation part, we computed the average translation of the x, y, and z direction. Since the points in the source point cloud are randomly selected on each precision iteration, different correspondences between the source and target point cloud are observed, leading to a slightly different transformation. The standard deviation of this translation that is calculated for each x, y and z element in the transformation matrix, gives us the precision and is shown in the following three figures, 13, 14, and 15.

The variance of the x-translation can be observed in Figure 13. As can be seen, the value of the variance of PCA is zero. This is because of the different steps PCA undergoes to achieve an affine transformation. The variance on the centroid position of the source point cloud will not change a lot if a few points a missing. ICP point-to-surface has a lower variance in x-translation than ICP point-to-point due to surface normal estimation. The advantage of the surface estimation makes the ICP point-to-surface approach more precise due to low changes of surfaces. In comparison to the results of the robustness is the variance of applying an SVD approach before an ICP point-to-point method worse than without a coarse registration approach. Solving the problem by a non-linear cost-function, such as a Huber-Loss function, will result in the worst precision. These benchmark results are only applicable for indoor environmental data, that is retrieved with a Microsoft Kinect Camera.

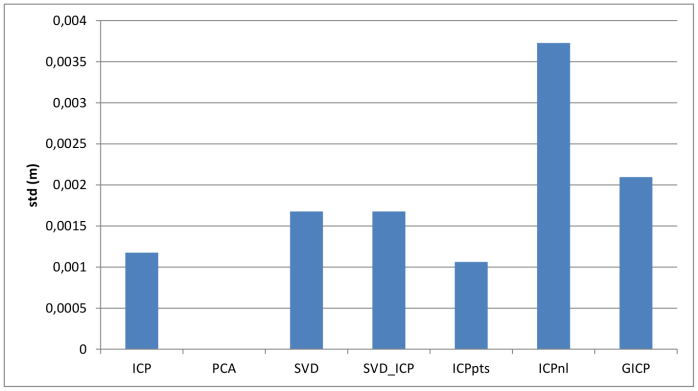


Figure 13. The x-axis represents the different methods and the y-axis represents the variance of the precision test in the x direction

When observing the variance of the translation in the y direction, a remarkable result for the non-linear approach can be seen in Figure 14. These results are much worse than in the x direction. This could be the result of setting the number of ICP iterations too low. As for the non-linear approach it is important to choose this number of iterations correctly because of the different minimization cost-function. In order to ensure a fair competition between the different algorithms we set the number of ICP iteration fixed to 40. As can be seen in Figure 11, each algorithm has reached a global minimum at 40 iterations. The other approaches have a similar result for the y direction as for the x direction.

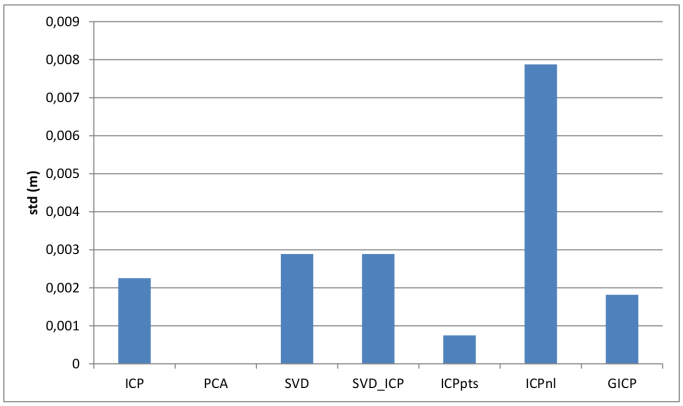


Figure 14. The x-axis represents the different methods and the y-axis represents the variance of the precision test in the y direction

The precision benchmark for the z direction gives better results than the x and y direction. Unlike the x and y directions we expected that the z direction, which represents the depth measurement, will give worse result due to noisy point clouds. The result of the variance of the z direction conclude that the precision of PCA is zero in all directions. This is because PCA translates the centered source point cloud against the centroid of the target point cloud and secondly, because PCA will not optimize the result. Thus, we can conclude that ICP point-to-surface has the best precision for the translation part.

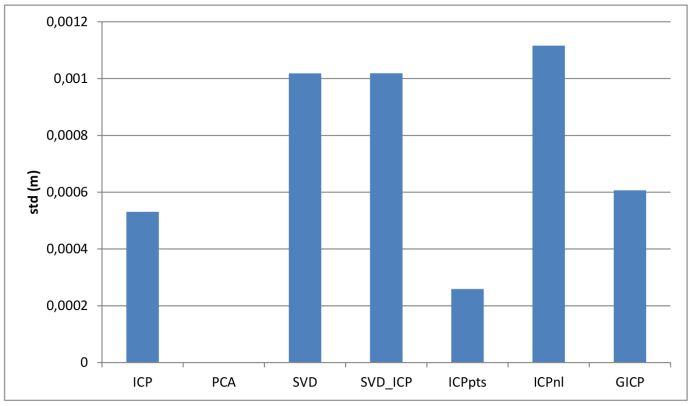


Figure 15. The x-axis represents the different methods and the y-axis represents the variance of the precision test in the z direction

The following figures shows the results of the precision for rotational part of the transformation. The 3×3 rotation matrix has been converted to Euler angles, in which each rotation is represented independently from each other by yaw, pitch

and roll. First, Figure 16 gives more insight to the variance of the different registration methods for the yaw rotation. The figure shows a remarkable difference for the PCA approach. This is because PCA observe the whole point cloud through the correlation between the different points by using the covariance matrix, while the ICP and SVD approaches will look for point correspondences. The variance in yaw direction is large due to the different subsamples, which will create point clouds where the density can change a lot in the direction of the smallest eigenvalue. This means that the probability of changing the direction of the largest eigenvector is large and thus the yaw rotation has a lower precision than the correspondence based approaches.

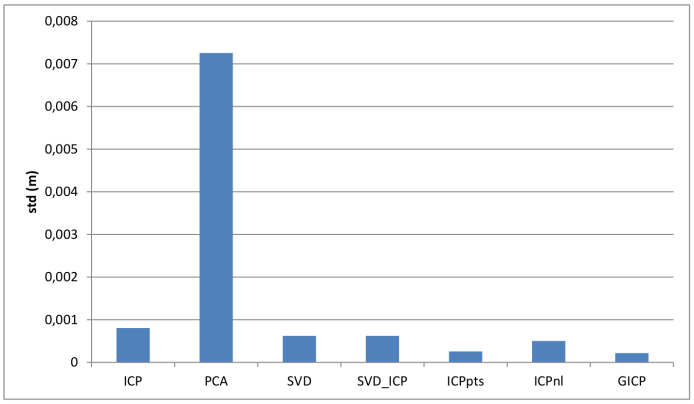


Figure 16. The x-axis represents the different methods and the y-axis represents the variance of the precision test in the yaw direction

To illustrate the precision of the transformation matrices after aligning with the different registration methods in the pitch direction. This result can be seen in Figure 17. The PCA method will perform more precisely in the pitch direction then the yaw direction. Secondly, the ICP point-to-surface approach will give the best results due to normal vector extension, which is a good parameter that is not changing a lot in the different subsamples of the source point cloud. Additionally, the variance of the method where the ICP approach is applied after a SVD is worse than the ICP point-to-point and the ICP point-to-surface methods.

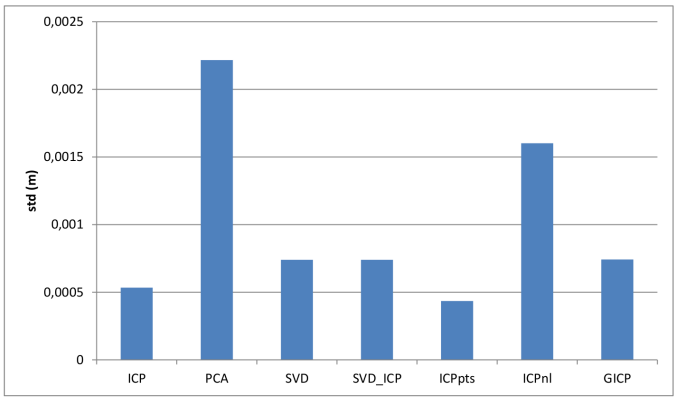


Figure 17. The x-axis represents the different methods and the y-axis represents the variance of the precision test in the pitch direction

The variances of the roll rotations are visualized in Figure 18. The algorithm that performs best is the ICP point-to-surface approach. Additionally, GICP performs better than ICP point-to-point method, the difference between these algorithm are negligible.

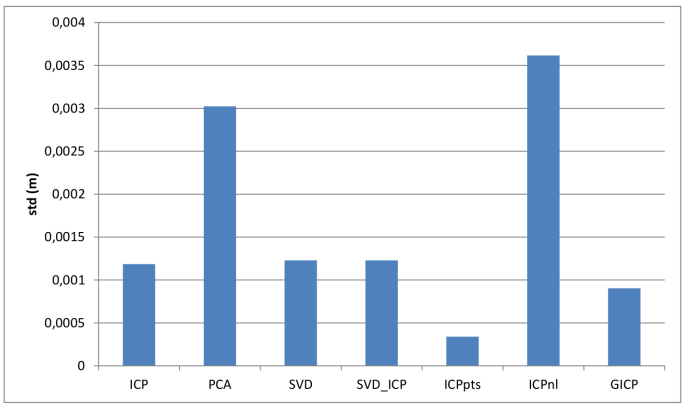


Figure 18. The x-axis represents the different methods and the y-axis represents the variance of the precision test in the roll direction

The different visualizations show that the result of the ICP point-to-surface method is the most rotation precise registration method. Followed by the GICP that has the best precision in yaw direction and the third in pitch. Due to the fact that the yaw direction is more valuable than the pitch, GICP is the second most precise algorithm based on rotational part of the transformation. The reason why yaw is more valuable than pitch is specific for this case where we want the most precise algorithm for a mobile robot SLAM application where the yaw rotation can change a lot in comparison with the pitch rotation. The ICP point-to-point algorithm results in the third most precise algorithm. This result is based on the rotational part of the transformation.

VI. Conclusion

This survey paper provides an overview of six different rigid 3D registration methods commonly used in robotics and computer vision. We discussed the mathematical foundations that are common to each of these algorithms and showed that each of them represents different approaches to solve a common least-squares optimization problem.

Finally, we compared the different methods with a critical view on their performance on a dataset, that was created with a Pioneer-3DX robot and a Microsoft Kinect Camera. To illustrate the performance, we quantified the robustness and the precision of the different registration methods. As result for the robustness we can conclude for this dataset that a combination of applying a ICP point-to-point method after an SVD method gives the minimum error based on 165 different point cloud pairs. On the other hand, the ICP point-to-surface is the most precise algorithm based on the rotational and translational part of the transformation after applying the precision benchmark test of this paper.

References

[1] B. Bellekens, V. Spruyt, and M. Weyn, “A Survey of Rigid 3D Pointcloud Registration Algorithms,” in AMBIENT 2014, The Fourth International Conference on Ambient Computing, Applications, Services and Technologies., 2014, pp. 8–13.

[2] R. B. Rusu, "Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments," KI - Künstliche Intelligenz, vol. 24, no. 4, Aug. 2010, pp. 345–348. [Online]. Available: http://link.springer.com/10.1007/s13218-010-0059-6

[3] R. A. Newcombe, A. J. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in 2011 10th IEEE International Symposium on Mixed and Augmented Reality. IEEE, Oct. 2011, pp. 127–136.

[4] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, Nov. 2013, pp. 2100–2106.

[5] S. Marden and J. Guivant, "Improving the Performance of ICP for Real-Time Applications using an Approximate Nearest Neighbour Search," 2012, pp. 3–5.

[6] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," P. S. Schenker, Ed., Apr. 1992, pp. 586–606.

[7] S. Fantoni, U. Castellani, and A. Fusiello, "Accurate and automatic alignment of range surfaces," in Proceedings - 2nd Joint 3DIM/3DPVT Conference: 3D Imaging, Modeling, Processing, Visualization and Transmission, 3DIMPVT 2012. IEEE, Oct. 2012, pp. 73–80.

[8] A. V. Segal, D. Haehnel, and S. Thrun, "Generalized-ICP," in Proceedings of Robotics: Science and Systems, Seattle, 2009, p. 8.

[9] J. Salvi, C. Matabosch, D. Fofi, and J. Forest, "A review of recent range image registration methods with accuracy evaluation," Image and Vision Computing, vol. 25, 2007, pp. 578–596.

[10] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images." IEEE Transactions on Medical Imaging, vol. 18, no. 8, Aug. 1999, pp. 712–21. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10534053

[11] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard, "An evaluation of the RGB-D SLAM system," Robotics and Automation (ICRA), 2012 IEEE International Conference, vol. 3, no. c, May 2012, pp. 1691–1696. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6225199

[12] J. Aulinas, Y. Petillot, J. Salvi, and X. Lladó, "The SLAM problem: a survey." CCIA, 2008, pp. 363—-371.

[13] P. F. I. N. D. E. Carrera, "MADRID RGB-D SLAM Author : Jorge García Bueno," no. October, 2011.

[14] K. Berger, S. Meister, R. Nair, and D. Kondermann, "A state of the art report on kinect sensor setups in computer vision," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8200 LNCS, 2013, pp. 257–272.

[15] J. Sprickerhof and A. Nüchter, "An Explicit Loop Closing Technique for 6D SLAM." ECMR, 2009, pp. 1–6. [Online]. Available: http://plum.eecs.jacobs-university.de/download/ecmr2009.pdf

[16] A. S. Huang and A. Bachrach, "Visual odometry and mapping for autonomous flight using an RGB-D camera," International Symposium on Robotics Research (ISRR), 2011, pp. 1–16.

[17] M. Ruhnke, L. Bo, D. Fox, and W. Burgard, "Compact RGBD Surface Models Based on Sparse Coding." AAAI, 2013.

[18] S. Savarese, "3D generic object categorization, localization and pose estimation," in 2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007, pp. 1–8.

[19] R. Shams, P. Sadeghi, R. Kennedy, and R. Hartley, "A survey of medical image registration on multicore and the GPU," pp. 50–60, 2010. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5438962

[20] D. Lee, H. Kim, and H. Myung, "Image feature-based real-time RGB-D 3D SLAM with GPU acceleration," Journal of Institute of Control, Robotics and Systems, vol. 19, no. Urai, 2013, pp. 457–461.

[21] W. R. Crum, "Non-rigid image registration: theory and practice," British Journal of Radiology, vol. 77, no. suppl_2, Dec. 2004, pp. S140–S153.

[22] J. Kay, "Introduction to Homogeneous Transformations & Robot Kinematics," Rowan University Computer Science Department, no. January, 2005, pp. 1–25.

[23] B. Draper, W. Yambor, and J. Beveridge, "Analyzing pca-based face recognition algorithms: Eigenvector selection and distance measures," Empirical Evaluation Methods in Computer Vision, Singapore, 2002, pp. 1–14.

[24] K. Low, "Linear least-squares optimization for point-to-plane icp surface registration," Tech. Rep. February, 2004.

[25] R. B. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL)," in IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China, May 2011.

[26] "ROS," 2015. [Online]. Available: http://www.ros.org/

[27] G. Grisetti, C. Stachniss, and W. Burgard, "Improved Techniques for Grid Mapping," Robotics, IEEE Transactions on Robotics, pp. 1–12.

[28] D. Fox, W. Burgard, F. Dellaert, and S. Thrun, "Monte Carlo Localization: Efficient Position Estimation for Mobile Robots," Aaai-99, no. Handschin 1970, 1999, pp. 343–349. [Online]. Available: http://dl.acm.org/citation.cfm?id=315322

Semantic Support for Tables using RDF Record Table

Mari Wigham[1], Hajo Rijgersberg[1], Martine de Vos[2], and Jan Top[1,2]
[1] Wageningen UR, Food and Biobased Research
Wageningen, The Netherlands
{firstname.secondname@wur.nl}

[2] VU University Amsterdam
Amsterdam, The Netherlands
{firstname.prefix.secondname@vu.nl}

***Abstract*—Tabular datasets are common in many domains, for example science and engineering. These are often not very well specified, and are therefore hard to understand and use. Semantic standards are available to express the meaning and context of the data. However, present standards have their limitations in expressing heterogeneous datasets with several types of measurements, missing data, and irregular structures. Such datasets are abundant in everyday life. We propose the RDF (Resource Description Framework) Record Table vocabulary for semantically modelling tabular data, as a supplement to the existing RDF Data Cube standard. RDF Record Table has a nested structure of records that contain self-describing observations, and is able to cope with irregular, missing and unexpected data. This allows it to escape the constraints of RDF Data Cube and to model complex data, such as that occurring in science and engineering. We demonstrate our Excel add-in for transforming data into the Record Table format. We propose a general approach to integrating tabular data in RDF, and confirm this approach by implementing integration support in the add-in and evaluating this in industrial use cases. This semantic support for tables helps researchers and data analysts to get the most out of available quantitative data with a minimum of effort.**

Keywords - semantics; table; spreadsheet; e-science, integration.

I. INTRODUCTION

Tabular data are common in many domains, for instance science and engineering. Tools to handle such data, such as spreadsheets, are extremely popular because of their flexibility and ease of use. However, this flexibility often leads to data being ambiguous or even incomprehensible, and their provenance being unknown [1][2][3]. The possibility to immediately proceed to the analysis and visualization of the data can have a negative effect on the quality of the actual data registration in terms of complete and systematic recording. Our work on introducing electronic lab notebooks in the multidisciplinary domain of food science has revealed many issues in data recording in the lab. Annotation of the data is often scarce and ambiguous due to the focus of researchers on the research itself rather than bookkeeping. In addition, large amounts of data are produced by automated measurement equipment in the lab. These devices tend to produce more systematic metadata, but linking data from different sources is as yet difficult and labor intensive. This makes finding, understanding and reusing the data very difficult [4]. As the amount of available data is exploding, it is essential to be able to efficiently locate and reuse existing datasets.

The traditional way to present tabular data is in tables on paper or on a screen. Rows and columns of cells make up their structure, and these cells are filled with simple data types such as numbers, strings or dates. In such a table, an individual recording shows up as a single value in one of the table cells. The associated header cell along the same column or row explains the meaning of this value, for example ‘m (kg)’ for mass measured in kilograms. In datasets found in practice, this header information is often ambiguous and incomplete. In fact, much of the information about the actual observation is frequently left out. This may even be done on purpose, in order to clean the data for presentation or processing. Tables also become more compact if all records contain the same quantities, the same unit of measure and have the same interpretation. In this way, the ‘bare’ numerical or string value in the table cells is separated from the metadata, and is directly visible for comparison and available for numerical computation. Researchers are trained in reading such tables and can usually interpret the meaning of the structure immediately. However, ambiguities in the structure can still arise, for example empty cells may be intended by the author to convey that the content of the previous cell should be repeated, but may cause confusion in a reader.

While the structure is usually easy to interpret, the frequently ambiguous and incomplete content of the headers gives readers more trouble. Abbreviations, ambiguous indications of quantities and units, language differences, jargon and typos all contribute to spreadsheets being frequently incomprehensible to all but the author. After time has passed, even the author may have trouble.

Interpreting such spreadsheets correctly is therefore hard enough for human beings, but next to impossible for a machine. This cuts off an enormous source of potential support for users. With all the computing power at their disposal, they are reduced to browsing through data files to find the one they need, and cutting and pasting data to combine it.

Fortunately, for the further exploitation of datasets, we are not bound to this traditional representation of a table. We can use richer representations to express more contextual information by using semantic technology. Many semantic methods have been developed over the last decades to express tabular datasets in a richer, more flexible manner. The W3C RDF (Resource Description Framework) standard provides a general, graph-based language for describing datasets [5]. RDF Data Cube is a prominent example of an RDF-based standard for expressing tabular datasets [6].

Representing datasets semantically has major advantages. Firstly, the meaning of the measurements is independent of the precise text in a spreadsheet, so that data can be found and understood regardless of typos, abbreviations, local terminology and even different languages. Secondly, the use of semantic concepts makes tables machine readable, meaning that they can be (semi-) automatically processed, from simple unit conversion up to complex computations. Finally, allowable numerical values and units can be defined, making it possible to check or clean the data. Moreover, semantic tables can be used as templates for future observations and experiments.

Initially, we proposed spreadsheet templates to stimulate systematic annotation of research data, but experience has shown that this restricts the creative and essentially unstructured character of scientific research. Therefore, a standard is necessary that facilitates annotation in a way that is flexible enough to accommodate researchers' needs.

Which requirements should a semantic standard meet to facilitate and stimulate structured annotation of tabular data? Firstly, it should be able to annotate the individual data elements, the content. For example, it should be possible to state that 'the mass of this sample is 2.95 grams', 'the city considered is Amsterdam', or 'this event occurred 5 minutes and 6.3 seconds later'. Good scientific recordings contain extensive information about each observation, for example on which object it has been measured, by which method and by whom. The annotation (metadata) of the individual data elements explains them and describes their provenance and relations. The keystone of semantics is the idea of an ontology, a sort of vocabulary that describes shared concepts and the relationships between them. A standard has to build on existing (domain) ontologies in order to facilitate shared understanding of the individual observations.

Secondly, a semantic standard for tabular data should make explicit the structure – the grouping together of scientific observations that collectively form a 'snapshot' of the world. The observations may be combined because they are generated in one experiment, using the same experimental protocol or by a single apparatus, or for a multitude of other reasons. A collection of snapshots, or *records*, is used to detect patterns, similarities or correlations.

This grouping is essential for correct interpretation of the data. Within one experiment, the structure of the records is often quite similar. However, when comprehensive recording of all possibly relevant effects is required, datasets can be less homogeneous and well-formed. This holds for

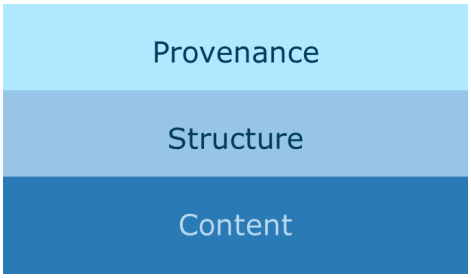


Figure 1: The three components of a semantic standard for tabular data

datasets that combine observations from different origins, in particular. Moreover, exact science typically deals with quantities having diverse scales, units and other specifications; values may be missing or occasionally additional measurements are available. Consider for example research that combines input from a number of labs around the world. Some of them have recorded the environmental temperature in degrees Fahrenheit and others in degrees Celsius. One lab has not measured temperature at all. Semantic standards should allow these variations and at the same time provide enough structure to preserve the meaning of the data.

Thirdly, a semantic standard for tabular data must make it possible to link to provenance information, to indicate where the data came from. Well-publicized cases of fraud in scientific research make the traceability of data a central concern to many research institutes. Fig.1 shows the three components of a semantic standard for tabular data.

Finally, the semantic standard must be flexible enough to accommodate the variations present in scientific data, and be implemented in tools already in use by researchers, in order to harmonize with their research work, rather than distracting from it.

In this paper, we discuss RDF Record Table, a format that is sufficiently rich and flexible to handle complex datasets, such as those often found in science and engineering. RDF Record Table was first introduced in [1]. In this paper, we will expand on the description of the model and discuss its benefits in more detail. In Section II, we first briefly describe existing approaches and tools for modelling tabular data in RDF. In Section III, we go into more detail on the RDF Data Cube vocabulary. This is a recommended W3C standard for multidimensional tables. To be able to handle more heterogeneous datasets, we propose RDF Record Table in Section IV, as a supplement to RDF Data Cube. RDF Record Table uses self-contained observations and recursive records. In Section V, we describe how we can reduce redundancy and include header information in the model by allowing cells to refer to other, similar cells. Section VI discusses the differences between RDF Data Cube and RDF Record Table, in particular with reference to specific challenges faced in scientific data. This is followed by a description of a first implementation for annotating and transforming spreadsheet data to RDF Record Table in Microsoft Excel in Section VII. We then discuss how RDF Record Table makes it easier to integrate

data in Section VIII. We describe our approach to data integration, and explain how this approach is implemented in the Rosanne add-in. The approach is validated in a number of use cases in Section IX. Finally, we conclude in Section X, also listing a number of open issues.

II. RELATED WORK

Many methods take the relational database approach when they convert tables or databases into an RDF-based representation [7]-[9]. They assume that a table consists of a header row defining variables, and other rows that contain strings or numbers representing the value of the variable in the same column. In general, they do not support more complex structures. All columns are translated into RDF properties of a single object. At this point, no other metadata is available than which is given in the header and data cells, the information implied by the table structure is lost. For simple data this may suffice, however, problems quickly arise with more complex data. For example, repeated measurements of the same property will simply produce triples with two different values, without the context that would allow an understanding of why the measurements are different (different time, different apparatus, etc.). The information contained in one column may also be necessary to correctly interpret the information in other columns. For example, based on the price information of two types of cheese, you might conclude that the cheese with the lower price is cheaper. However, the amount of cheese, recorded in a different column, could be completely different. Tables also frequently contain information that is not a property of the same single object, for example, the temperature of the room in which the density of the sample was measured. Finally, removing all indication of table and cells by converting only the data to triples, removes the ability to convey provenance information about the data.

A richer format is defined by the RDF Data Cube vocabulary [6], a recommended W3C standard. This vocabulary has been developed in the context of statistical data in social sciences and policy studies, but is also being applied in other areas [10]. Information about the meaning of the data is expressed by linking to concepts from other ontologies, most typically the SDMX vocabulary [11]. These individual data observations are stored in a multi-dimensional hypercube structure to preserve the relationship between the measured values and the dimensions along which they vary, such as time, location, gender, etc. Metadata can be linked to individual observations, parts of datasets, or whole datasets.

There are various tools that have been developed for converting tabular data into RDF in general, or RDF Data Cube in particular. The EU CODE project [12] developed the CODE platform, which extracts tabular data from PDFs, csv-based documents or existing RDF repositories and converts it to RDF Data Cubes. These cubes can then be visualized. The Tabels (sic) project [13] attempts to discover the data structures in tabular data and transform these to RDF Data Cube. TabLinker [14] and RightField [15] assist the user in annotating their numerical data, which is then converted to RDF Data Cube, in the case of TabLinker. CSV2Data Cube [16] helps the user to configure dimensions and attributes from their CSV file. It then transforms the data to RDF Data Cube. The OpenCube toolkit [17] [17]from the EU OpenCube project [18] allows relational data and csv/tsv files to be converted to RDF Data Cube. These cubes can then be visualized and also submitted to statistical analysis. Tools for visualization, slicing and validation of RDF Data Cube fall outside the scope of this work.

The available tools are mostly directed at the domain of statistical data and, with the exception of RightField (which does not handle table structure), appear to be limited to simple table structures. Statistical data is, as a rule, much more uniform and regular than scientific or engineering data, which can have quite complex table structures.

All of these tools are separate to the tools that are usually used by researchers in the course of their work (with the exception of RightField, which generates templates that are used in Excel). This requires researchers to interrupt their workflow in order to carry out data documentation. This can be a barrier for researchers.

We have found one incidence of related work on representing more complex, irregular data in RDF [19] investigated linked Data Cubes for clinical data. Some of the difficulties they experienced could be solved by augmenting RDF Data Cube with constructs from other vocabularies, others remained unsolved.

Whereas RDF Data Cube and other standards define the structure and context of tabular data, they are not intended for expressing provenance of data on the web. However, they do provide identifiers for the data, which can be linked to a description of the provenance of that data. For that purpose, additional vocabularies have been developed. The W3C-standard PROV is becoming increasingly popular for this purpose [21]. It describes the origins of any type of data, helping the user to evaluate how appropriate and trustworthy the data is for a particular use. PROV basically says that a `prov:Agent` performs a `prov:Activity`, in which he uses or generates a `prov:Entity`. Tables, records, slices and individual measurements can all be seen as subclasses of `prov:Entity`. The previously defined Dublin Core Terms [13] vocabulary complements the PROV model with detailed concepts about publications and authorship.

We wish to develop a standard for tabular data that can handle the sort of complex, irregular data that is found in many practical situations. This standard will be able to be linked to the PROV standard and will be implemented in tools that researchers already use in their daily work.

III. RDF DATA CUBE

RDF Data Cube organizes observations as multidimensional datasets. Each observation is a point in n-dimensional space, defined by the associated values of the *dimensions*. Typical dimensions in RDF Data Cube are ‘time’, ‘area’ and ‘gender’. Each observation contains one or more *measures*, for example ‘life expectancy = 83.5’.

Table I: Life expectancy data in different regions over time

	2004-2006		2005-2007		2006-2008	
	Male	***Female***	***Male***	***Female***	***Male***	***Female***
Newport	76.7	80.7	77.1	80.9	77.0	81.5
Cardiff	78.7	83.3	78.6	83.7	78.7	83.4
Monmouthshire	76.6	81.3	76.5	81.5	76.6	81.7
Merthyr Tydfil	75.5	79.1	75.5	79.4	74.9	79.6

Observations, measures and dimensions can have *attributes* that provide additional information about them, for example the unit of measure used. A separate section of an RDF Data Cube defines its *structure*; this section can be used as a template for future observations. Another section gives information for external reference to the entire dataset.

In its normalized form, each observation in a data cube contains all its dimensional values. One way to reduce redundancy is by moving shared attributes to the structure definition section. Further reduction can be obtained by introducing 'slices'. A slice is a lower-dimensional representation, which also serves as a proposed interpretation of the dataset. Moreover, one can refer to a slice as an independent entity. This allows easy access to predefined views of the data.

In order to group together observations that do not fulfil the requirements of a slice, the concept of ObservationGroup is defined in RDF Data Cube. This allows any observations, even from different datasets, to be grouped together.

Table I shows the example table that the RDF Data Cube definition uses to explain the vocabulary [6]. The full RDF Data Cube model of Table I is available for viewing at [6].

The RDF Data Cube vocabulary is very well suited for modelling well-formed, complete datasets such as are produced by statistics offices. Software tools are available to provide useful views of the data. However, these advantages are the result of some restrictions on the data. RDF Data Cube is intended for describing 'well-formed' datasets. As a result, several constraints are placed on the data, for example that each observation must have a value for every measure. For example, if for one measurement in the example it is not known whether this person is a man or a woman, then this data point cannot be included in the model. Another assumption is that the multidimensional structure is a regular (hyper)cube, not permitting rows with varying length for a single dimension. If we know the standard deviation of the life expectancy value for Cardiff and Newport, but not the other regions, we cannot add this to the above Table I.

RDF Data Cube has two alternative ways to handle datasets with more than one measure, which cannot be used simultaneously. In the *multiple measures* approach one observation can contain more than one measured quantity. However, all quantities must have the same attributes, for example, the same type and unit of measure. This rules out this approach for many exact science applications. The second approach restricts observations to having a single measured value. It allows a dataset to carry multiple measures by adding an extra dimension, a measure dimension. This turns a measured value into a kind of semi-dimension.

Another characteristic of RDF Data Cube is that it makes extensive use of properties (rather than classes) as its main organizing mechanism. The design introduces many different types of properties. It is questionable whether these different properties are needed to express the meaning of the data. They make the design of a model rather complex.

As datasets, slices, ObservationGroups and observations all have unique identifiers in RDF Data Cube, they can all be referred to by a provenance model, enabling the provenance of the data to be traced.

RDF Data Cube is the only semantic standard currently available which explicitly and thoroughly models the structure of tabular data.

IV. RDF RECORD TABLE

Experience with researchers over the past ten years has confronted us with many different datasets. Many of them are contained in spreadsheets and data analysis tools such as Matlab [22], SPSS [23] and R[24]. Inspired by other initiatives to annotate datasets using RDF, we have devised an approach that can work in the tools commonly used by researchers and at the same time support rich annotation. This approach has at its heart a model for tabular data called RDF Record Table.

The RDF Record Table vocabulary is intended for recording original and processed data across all domains, including science and engineering in particular. It is based on the observation that the common two-dimensional table in reports and spreadsheets is a restricted representation of a more general graph-based table model. A human reader of a table in a report or spreadsheet implicitly combines his or her interpretation of the text in individual table cells with the visual inspection of the table layout (topography, coloring, typesetting, etc.). This forces authors of tables to express two types of information in a two-dimensional format that it is not ideally suited for, viz., (i) nesting of records and (ii) describing metadata. In this section we show how the RDF Record Table model deals with these issues by supporting recursive nesting of records and by enriching data elements with metadata. In the next section, we will show how the model supports sharing of metadata between multiple data elements. RDF Record Table models the structure of tables in terms of cells and records (see Fig. 1, using rec: as a prefix for the RDF Record Table namespace). A *cell* contains a statement about an entity or the property of an entity, such as 'the temperature of this object measured by a pt-sensor is 36.5C' or 'this milk sample is from batch 20140612YTU'. A *record* combines cells in a group, thus conveying the assumption that in some way the observations are related - in time, location, subject, conditions, or in another way. This assumption can be made when setting up a new experiment, but also when existing data are combined. It is similar to the ObservationGroup concept in RDF DataCube, but in RDF Record Table it is a core element rather than an optional extra. Scientific and

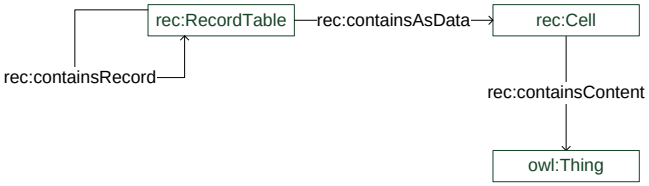


Figure 2: Basic RDF Record Table schema

engineering data are grouped and regrouped continuously to investigate hypothesized correlations and causalities. We submit therefore that the structure of the data should be flexible and based around the groupings chosen by the researcher. To express composite structures, in RDF Record Table any record can recursively contain sub-records, which again are of the type `rec:RecordTable`. This means that we do not make a distinction between the concept table and the concept record. After all, both are simply groupings of data. For example, an experiment may observe multiple samples at one fixed temperature. For each sample its viscosity, composition and mass are measured over time. This means that the entire dataset consists of a Record Table that at its highest level contains (i) the observed temperature and (ii) a sub-record for each sample. Each sub-record in turn contains the sample identifier and sub-records that describe viscosity, composition and mass for that sample measured at a point in time. In the most explicit form, all sub-records are expanded into non-nested records. In this example, the top level Record Table only contains sub-records, each of them stating the observed temperature, time point, sample id and the other measured properties.

RDF Record Table is shown in Fig. 2. In Turtle format, it is defined as follows.

```
rec:RecordTable  a   rdfs:Class ;
        rdfs:subClassOf  prov:Entity .

rec:Cell   a   rdfs:Class ;
        rdfs:subClassOf  prov:Entity .

rec:containsAsData  a   owl:ObjectProperty ;
        rdfs:domain  rec:RecordTable ;
        rdfs:range   rec:Cell .

rec:containsContent  a  owl:ObjectProperty ;
        rdfs:domain  rec:Cell ;
        rdfs:range   owl:Thing .
```

The next question is how the cells in the nested records can contain the actual observed values in such a way that they can be properly understood both by human users and machines. From the inspection of many tables used in practice, we see that two types of observations frequently occur: (i) *identified entities* and (ii) *properties measured on a scale*. Examples of *identified entities* are ‘sample XY876b’, ‘Newport’ and ‘Peter’. Quantities such as ‘length’, ‘mass’, and ‘temperature’ are examples of

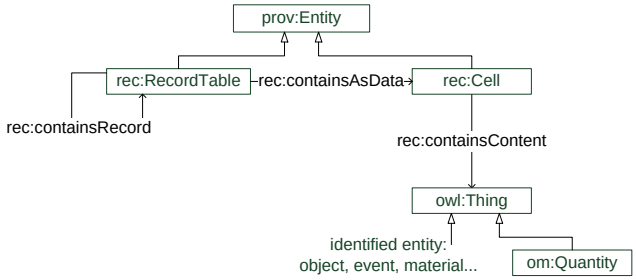


Figure 3: RDF Record Table expressing domain and provenance information

properties measured on a scale. These two types are not formally part of the RDF Record Table model, which allows any ‘Thing’ to be in a cell. However, we propose this distinction as a best practice that works in many cases. Fig. 3 shows how quantities and identified entities fit into the RDF Record Table model.

In traditional tables, identified entities are typically represented by a human readable identifier, and an explanation of the entity type in the associated header cell. For example, ‘Peter’ is a unique name for an entity of type ‘Author’. RDF Record Table uses externally available domain ontologies to express all that is needed to know about such an entity by pointing to the respective instance in an RDFS/OWL schema. For modelling Table I we have chosen to view instances of ‘Area’ and ‘Period’, such as 2004-2006, as identified entities since they are not supposed to be read as nominal or even numerical values.

For the other type of observation, a *property measured on a scale*, RDF Record Table uses ontologies that define quantitative or qualitative values defined on a scale, possibly with units of measure. In Table I, ‘Sex’ and ‘Life Expectancy’ are typical measured properties, one on a nominal scale and the other on a rational scale with unit ‘Year’. In our work we use OM (Ontology of units of Measure and related concepts) [25] for expressing quantitative measurements. OM contains a large number of quantities and units of measure suited to scientific and engineering datasets. It also provides the necessary properties for linking the quantities, domain concepts and units. However, other ontologies such as QUDT [26] and SDMX [11] can be used equally well. The measured quantities can be properties of the observed entities in the table, but do not need to be related to anything specific. For example, in Table I, the life expectancy measured is that of people in the associated geographical region. On the other hand, ‘the time of day’ is usually not connected to a specific entity (except for example to a ‘time zone’ that relates to a geographical area).

This division into *identified entities* and *properties measured on a scale* is highly useful, as it relates to the type of data handling that is typically applied to data from each category. Measured properties are usually subject to numerical processing, and require units of measure. Identified entities on the other hand may be used as identifiers on which, for example, different tables can be

joined. While this distinction can assist processing, it does not limit it – for example, tables may also be joined on numerical values, if the user wishes.

Finally, by making `rec:RecordTable` and `rec:Cell` subclasses of `prov:Entity` we ensure that all provenance information can be expressed for individual measurements, records and tables. For example, the relation `prov:wasDerivedFrom` between two cells tells us that the quantity in one cell depends on the value of the quantity in the other cell.

To illustrate the use of the RDF Record Table format, we show how the cells with values 76.7 and 83.3 in Table I are contained in the table. We see that the first level of nesting defines four records (:o1, :o2, :o3, :o4), one for each region. We use the ontology for geographic areas (as *identified entities*) that was also used in the RDF Data Cube example [6]. The next level specifies the three time periods, again using instances that were also used in the data cube example. At the third level of sub-records, we register two properties measured on a scale, viz., 'sex' and 'life expectancy'. For indicating the variable 'sex', we use an sdmx-code, as in the data cube; to illustrate the use of OM [25], we use the concept `om:Duration` from that ontology to describe 'life expectancy'. The value of a quantity in OM is of the type `om:Measure`, which is a combination of a numerical value and a unit (or scale).

```
:dataset1  a  rec:RecordTable ;
   rec:containsRecord :o1 , :o2 , :o3 , :o4 .

:o1  a   rec:RecordTable ;
   rec:containsAsData :cell_newport ;
   rec:containsRecord :o11 , :o12 , :o13 .


:cell_newport  a     rec:Cell ;
    rec:containsContent  ex-geo:newport_00pr .


:o11  a   rec:RecordTable ;
   rec:containsAsData :cell_period_2004_2006 ;
   rec:containsRecord :o111 , :o112 .

:o111  a  rec:RecordTable ;
   rec:containsAsData  :cell_sex-M ,
       :cell_lifeExpectancy_76_7YR .

:cell_sex-M  a rec:Cell ;
   rec:containsContent  sdmx-code:sex-M .

:cell_lifeExpectancy_76_7YR  a rec:Cell ;
   rec:containsContent :lifeExpectancy_76_7YR ;

:lifeExpectancy_76_7YR a om:Duration ;
   om:value :_76_7YR .

:_76_7YR  a  om:Measure ;
   om:numerical_value "76.7"^^xsd:string ;
   om:unit_of_measure_or_measurement_scale om:year
.

...
```

```
:o2  a   rec:RecordTable ;
   rec:containsAsData :cell_cardiff  ;
   rec:containsRecord :o21 , :o22 , :o23 .

:cell_cardiff  a   rec:Cell ;
   rec:containsContent  ex-geo:cardiff_00pt.


:o21  a   rec:RecordTable ;
   rec:containsAsData :cell_period_2004_2006 ;
   rec:containsRecord :o211 , :o212 .

...

:o212  a  rec:RecordTable ;
   rec:containsAsData :cell_sex-F ,
       :lifeExpectancy_83_3YR  .

:cell_sex-F  a   :Cell ;
    rec:containsContent  sdmx-code:sex-F .

:cell_lifeExpectancy_83_3YR   a  rec:Cell ;
   rec:containsContent :lifeExpectancy_83_3YR .


:lifeExpectancy_83_3YR a om:Duration ;
   om:value :_83_3YR .

:_83_3YR  a  om:Measure ;
   om:numerical_value "83.3"^^xsd:string ;
   om:unit_of_measure_or_measurement_scale  om:year
.
```

To show the flexibility of the RDF Record Table model, we now show how a completely different type of measurement can be added to the above definitions, without changing anything in the previously modelled records and cells. In Table I, we add 'the measured average weight of the inhabitants of this region' to an existing record (`:o341`) using the OM quantity `om:mass`. In addition, we can switch to a value for 'life expectancy' measured in months rather than years for one single observation (74.9 years). The result is as follows:

```
:o431   a   :RecordTable ;
    rec:containsAsData  :cell_sex-M ,
        :cell_lifeExpectancy_898MONTH ,
        :cell_mass_71kg .

:cell_lifeExpectancy_898MONTH  a  rec:Cell ;
    rec:containsContent :lifeExpectancy_898MONTH .

:lifeExpectancy_898MONTH  a   m:Duration ;
    om:value  :_898MONTH .

:_898MONTH  a    om:Measure ;
    om:numerical_value  "898"^^xsd:string ;
    om:unit_of_measure_or_measurement_scale
                om:Month .

:cell_mass_71kg  a    rec:Cell ;
    rec:containsContent  :mass_71_kg .

:mass_71_kg  a  om:Mass ; om:value  :_71_kg .

:_71_kg  a  om:Measure ;
    om:numerical_value  "71"^^xsd:string ;
```

```
om:unit_of_measure_or_measurement_scale
            om:kilogram .
```

Note that so far we assume that each cell is entirely self-describing; it contains all the necessary information to know what kind of data it represents. Where data cells are similar, it is possible to use one description for many cells. We will discuss this in the next section.

V. HANDLING SIMILAR DATA ELEMENTS IN RDF RECORD TABLE

As stated before, traditional two-dimensional representations of tables express descriptive information about the data in the table headers. In their most basic form they form a single row at the top of a table, but much more complicated header structures occur commonly. Each header cell covers a range of data cells, typically shown in the column under the respective header.

In practice the distinction between header items and data items is not always clear. For example, in Table II, it is possible to view the top three rows and the left column as headers. In fact, only "Life Expectancy", "Period", "Area" and "Sex" are true headers, as they only supply descriptive information about the data. The other 'header' cells, such as "Male", and "2004-2006", actually supply different data values for one data type. This style of table, where the 'header' contains data values, is often called a 'pivoted table', as it can be produced by pivoting a 'flat table', where the header only contains descriptions of the data. Pivoted tables can give extra insight into data by grouping together data for which one field always has the same value, for example all data relating to "Cardiff". Any RDF Data Cube with more than one dimension is in the style of a pivoted table, the effect of choosing between a dimension and a measure is to select the measurements on which the data will be pivoted. A pivoted table can be 'unpivoted' by adding the data elements from the header to each record.

The RDF Record Table model defined in the previous section assumes that all data in the table cells are entirely self-descriptive. Each data element describes what kind of data it represents. For example, 't_{final} = 42 sec' expressed using OM concepts says that an activity has ended after 42 seconds. Traditional tables in reports and spreadsheets usually summarize this explaining information 't_{final} (sec)' in a table header, separately from the numerical '42', to make the table readable for humans and fit for numerical analysis. In RDF Record Table, in principle, we can do without such headers, as all this information is available in the data cells; in the above example the data cell would be linked to the concepts 'time' and 'seconds'. In the case of a pivoted table, the 'header' information is simply another data value

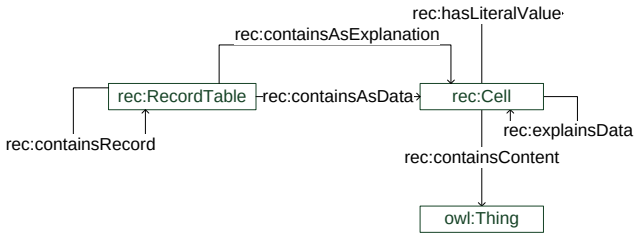


Figure 4: RDF Record Table with 'header' cells.

in the Record Table, with a nested Record Table containing the information that falls under the 'header'.

In practice, many data items in a single experiment are similar in some way – they refer to the same type of parameter, or play the same role in an experiment. We submit that the way that traditional tables express this, is inherently limited due their two-dimensional character. In RDF Record Table we look at table header cells in a generalized manner, independent of their usual two-dimensional representations. We assume that a 'header' cell explains a set of similar data items. It provides metadata that is not expressed by the individual data items.

There can be four reasons to put this information in header cells rather than in self-contained data cells. First, header cells can specify the type of measurement without giving actually observed values; they act as a prescriptive template for an experiment or for data analysis. Second, using header items for metadata is a way to remove redundancy and to achieve a significant reduction in the physical size of a dataset. For example, suppose that the temperature of an object has been measured over time. We can state in a header item that we have measured temperature, measured in kelvin, on a given object. In the corresponding data items we only have to state the numerical values. This takes much less space, but still allows us to regenerate the self-contained values for all data elements (if we know how to link the numbers to the reconstructed instances). Third, cells of different types may play the same role in a table. For example, a column containing numerical measurements may also include the entry 'measurement failed'. This cell clearly has a different type, but should still be grouped under the column – it plays the same role as the other measurements. Finally, the header cell itself may contain additional, informal information. For example, the header cell 'Area' may have contained the string 'Area, as per 2001 boundaries", or "Life Expectancy" may have been written using the Dutch word "Levensverwachting". While ideally this sort of information would also be modelled semantically, in practice this is not the case. If the header cell is modelled separately to the data cells, then its text content can be preserved exactly as it was, keeping any informal information and also making the table representation more familiar to the user.

Table II: Example extended with headers

Life Expectancy (years)							
Period		***2004-2006***		***2005-2007***		***2006-2008***	
Area	***Sex***	***Male***	***Female***	***Male***	***-***	***Male***	***Female***
Newport		76.7	80.7	77.1	80.9	77.0	81.5
Cardiff		78.7	83.3	-	83.7	78.7	83.4
Monmouthshire		76.6	81.3	76.5	-	76.6	81.7
Merthyr Tydfil		75.5	79.1	75.5	79.4	74.9	79.6

In RDF Record Table we do not define different types of cells for data and metadata since their internal structure is the same. Instead, we use the property `rec:explainsData` to have some cells act as metadata cells (see Fig. 4). Such an explaining cell contains an instance (or class) of a phenomenon or a quantity, acting as a template for other cells. The associated data cell in that case only needs to provide a numerical or string value (through `rec:hasLiteralValue`), all other information is derived from the explaining cell.

For example, in Table II, one header field states that we have observed 'Life Expectancy' in 'years'. We construct a cell in the RDF Record Table translation that contains an instance of `om:Duration` with unit 'year' but no numerical value assigned to it. The cell also preserves the string in the original table header by storing 'Life Expectancy (years)' as a literal value. The associated cells only have to provide a numerical value for each measurement and a reference to this header item. Or in another case, the header item 'Area' states that we have observed entities of type `admingeo:UnitaryAuthority`, and the data items express specific areas such as 'Monmouthshire', only using a literal string. When unpacking this compacted version to a fully self-describing model, the software has to match the string 'Monmouthshire' with instances of `admingeo:UnitaryAuthority` to find the proper instance `ex-geo:Monmouthshire`.

Suppose that no interpretation whatsoever is possible given a specific traditional table, except for its structure. In that case, we can translate this table directly into an RDF Record Table with only literal values, using only the property `rec:hasLiteralValue` for both header items and data items. If we can also map the rows and columns directly to records, data cells and explaining cells, this would be the least semantically rich representation possible. Once translated to RDF Record Table, we can process this information and possibly add more semantics to it. RDF Record Table allows each data cell to have its own specification, overruling the information in the header item.

In the previous section we have shown an example in which a single cell measured 'life expectancy' in 'months', whereas all others were measured in 'years'. In that case, we could have used an RDF Record Table model with an explaining cell that states that in principle all values are in years. The single cell that uses 'months' as a unit overrules this general statement for that specific cell.

Regenerating an RDF Record Table with semantically self-contained data items is possible if we know how to relate the information in partially specified data cells to the associated explaining items. When using for example the OM quantity `om:Duration` with unit `om:year` in a header cell, we know that a numerical value in an associated data cell specifies the 'numerical value' property of the Measure of this quantity. This is the type of interpretation that readers of tables on paper make all the time, but is not obvious for automatic processing. This knowledge has to be incorporated in the software that unpacks a model.

Fig. 4 shows the extended schema for RDF Record Table, including header information using the property `rec:explainsData`. It shows that the class `rec:Cell` can play the role of either header cell or data cell. As a matter of fact, it is possible to use a data cell as a header item. This means that other data items use this particular item to provide their type and other information, while the data items simply provide a numerical or string value.

In Fig. 4, we can see that records can directly indicate which cells play a role as header cells using the property `rec:containsAsExplanation`. For Table II such a listing of explanatory cells would be modelled as follows:

```
:dataset1  a  rec:RecordTable ;
   rec:containsRecord :o1 , :o2 , :o3 , :o4 ;
   rec:containsAsExplanation  :cell_sex ,
        :cell_lifeExpectancy_YR ,
        :cell_period_from_yr_to_yr
        :cell_geographicalArea .
```

The cell explaining the Life Expectancy measurements then refers to the data cells containing those measurements:

```
:cell_lifeExpectancy_YR  a  rec:Cell ;
   rec:explainsData
        :cell_lifeExpectancy_83_3YR ,
        :cell_lifeExpectancy_76_7YR ,
        :cell_lifeExpectancy_898MONTH ;
        :hasLiteralValue  "Life expectancy
(years)"^^xsd:string .
```

Although for reasons of clarity it is attractive to write the explaining cells as a 'header' at the first level of nesting, there is no formal need to do so. Such cells can be placed anywhere in an RDF Record Table model. This is useful when merging data from different sources, where the top level table is not known upfront.

Finally, we note that the property `rec:explainsAsData` is redundant if the respective cells are already used to explain data cells through the property `rec:explainsData`. However, this property can be used

to express the structure of an otherwise empty table, which then can serve as a template for new observations or analyses.

VI. Differences between RDF Data Cube and RDF Record Table

The most salient difference between RDF Data Cube and OQR Record Table is the fact that RDF Data Cube sees complex datasets as n-dimensional hypercubes, whereas RDF Record Tables are defined recursively via nesting. This use of n-dimensional hypercubes in RDF Data Cube has a profound impact on the type of data it can be used to model. RDF Data Cube expects a certain kind of dataset:

"At the heart of a statistical dataset is a set of observed values **organized along a group of dimensions**, together with associated metadata" [6]

RDF Data Cube also requires that cubes be well-formed, which requires, among other things, that there be no missing data, and that all measures and dimensions are required for all observations. In short, Data Cube expects a uniformly formed and filled cube, with no extra granularity in some areas that is missing in others, and no extra observations. This is often not the case, particularly when data from different sources is integrated together. In these situations, the integrity constraints can cause problems:

"Some specialised data cubes do not satisfy the integrity constraints, specifying that every qb:DataStructureDefinition must include at least one declared measure (IC-3), that only attributes may be optional (IC-6) and that each individual qb:Observation must have a value for every declared measure (IC-14). These constraints are too restrictive for our Nutrition data cube where the presence or ab-sence of a value for a particular category of food varies according to the subject's diet. This is a concern for survey questionnaires using previously entered values to determine if a field on a form should be mandatory filled." [19]

The authors of the above-mentioned paper specifically note the difference in their data from that commonly used in RDF Data Cube.

"The LCDC (Linked Clinical Data Cubes) use of the RDF Data Cube vocabulary is different from the more common use cases [10] primarily because of the unreliable, disparate and longitudinal nature of clinical data" [19]

RDF Record Table, on the other hand, has been specifically developed for researchers and their quantitative data, with extensive input from real-life research data. This data, like the data used by [19], is often far more irregular than most statistical data. In RDF Record Table any record can contain an arbitrary set of measurements, with different types and sub-records. Missing values or varying units of measure or other attributes within a single dataset are no problem. We do not demand completeness or regularity of the data, in the sense that a record can contain any set of entities and properties. This better reflects the reality of datasets in science and engineering, in particular, when datasets from different sources are combined. It can be argued that such datasets can be modelled in RDF Data Cube simply by violating the integrity constraints. However, this is a bad approach to using a standard, and can lead to interoperability problems between tools developed for the standard.

The second major distinction between the two approaches is that RDF Data Cube distinguishes between dimensions and measures, whereas OQR Record Table does not make a priori assumptions about the roles of individual observations. We consider making such decisions to be the task of the data analyst.

We will now further discuss the differences between RDF Data Cube and RDF Record Table in the context of specific challenges that are faced in annotating and integrating real-life data.

1) Missing data

According to the integrity constraints for RDF Data Cube, all data must be present. Naturally, even in the well-planned world of statistics bureaus, data is sometimes missing. There is no solution in the RDF Data Cube standard for this. However, the 'attribute' concept, which allows metadata to be attached to an observation, is a natural way to indicate missing data. In [10], a simple Boolean attribute is used to indicate when data is missing. It is of course then necessary that tools using the data are aware of this solution and can process it correctly.

When two tables are integrated together, there can be missing data even though both original tables were complete. For example, if one file contains the mass measurements for all dairy products, and another file contains viscosity measurements for all liquid dairy products, then when the two files are integrated together, there will be missing data for the solid dairy products. In order to cope with this in Data Cube, observations would have to be generated for these products, and then marked as 'missing data'.

In Record Table, there is no constraint requiring data to be present. Therefore, in the event of missing data, the cell can simply be omitted from the record. This fits better with the ethos of RDF.

Table III: Example of an irregularly nested table

			Installed Capacity (in 10 MW)		Operating Water Level (meter)	
Hydropower Stations	Budget (in hundred millions of CNY)	Completion Year	*Realisation level*		*Realisation level*	
			2003 Plan	*Upon Completion*	*2003 Plan*	*Upon Completion*
Twenty-Second	37.76		60	60	519	519
Twenty-First			15	15.5	533	533
Twentieth	79.37	2009	150	175	602	602

2) Unexpected data

In a table, a column can often contain an unexpected value. For example, a column of numerical measurements would be expected to contain values such as "1.34, 22452E-10". However, it is perfectly reasonable that a researcher may note down unexpected values: "<20, negligible, ~5". These are results that a human reader will be perfectly capable of processing when they appear in a table, but that can confuse a software tool.

In RDF Data Cube the ranges of the dimensions and measures are set. A column that expects to contain decimals, cannot therefore contain exceptions such as we name here. An option would be to store the value in an attribute, however, we regard the storing of data in a metadata field as highly inadvisable.

In RDF Record Table, the role of a cell is separate to its type. A cell containing content with type string, for example, 'negligible', can therefore be linked via the relationship `rec:explainsData` to a header containing content with type Mass. This allows the information that the mass is negligible to be stored with the correct data type, so it can be excluded from numerical processing such as aggregation. At the same time the role of the information is clear, which allows the value to be displayed along with the other mass measurements.

3) "Irregular" nesting

RDF Data Cube demands that every observation has a value for each dimension:

"Every `qb:Observation` has a value for each dimension declared in its associated `qb:DataStructure-Definition.`"[6]

This means that RDF Data Cube cannot model any tables that do not fulfil this requirement. Regularly nested tables, in which for each observation there is a value for each dimension, can be modelled without problems. However, tables are often partially or irregularly nested. The observations in these tables then do not have a value for each dimension. These data values are not missing as such, the dimension is simply not applicable for part of the data. Table III is a real-life example of such a table.

It is perfectly logical that this information about the constructed dams is stored in one table. However, this table cannot be modelled as one Data Cube, as the Completion Year and Budget observations do not have values for each value of the Realisation Level dimension, because their value is not affected by the Realisation Level. It would have to be split into two Data Cubes, one with Dam name as Dimension, and Completion Year and Budget as Measures (Table IV); and one with Dam name and Realisation Level as Dimensions, and Installed Capacity and Operating Water Level as Measures (Table V). Alternatively, the Budget and Completion Year data could be repeated for each Realisation Level, but this creates the misleading impression that there is a relationship between these data and the Realisation Level.

Either approach requires either a fairly advanced level of understanding from the user, or quite intelligent processing from the data input tool. Breaking up the table into two Data Cubes also loses the implicit relationship between the data, which must then be indicated in metadata or by grouping the Data Cubes in an ObservationGroup. An alternative solution, namely using the void:subset relationship to indicate a link between Data Cubes, was used by [19]. This underlines the need for this sort of nesting in real-life data.

In RDF Record Table, the concepts of Dimension and Measure do not exist. The table can simply be annotated as it stands, and the nesting of Record Tables allows the extra information on Realisation Level to be added in to only the relevant portions of the table. The data is kept together, and the original structure (with all its implicit information) is retained, without need for additional constructs such as ObservationGroup.

4) Multiple measures

In the above example, one table had two Measures – Installed Capacity and Operating Water Level. As explained in the section on RDF Data Cube, in such a situation the user must choose between modelling these with multiple measures, or with a measure dimension. For many situations the choice made may not matter in practice; however, the choice must always be made. For novice users

Table IV: The first of the two tables into which Table III must be split when modelled in Data Cube

Hydropower Stations	Budget (in hundred millions of CNY)	Completion Year
Twenty-Second	37.76	
Twenty-First		
Twentieth	79.37	2009

Table V: The second of the two tables into which Table III must be split when modelled in Data Cube

Installed Capacity (in 10 MW)			Operating Water Level (meter)	
Hydropower Stations	*Realisation level*		*Realisation level*	
	2003	*Upon Completion*	*2003*	*Upon Completion*
Twenty-First	60	60	519	519
Twenty-Second	15	15.5	533	533
Twentieth	150	175	602	602

this can be confusing. When integrating two tables, one of which uses multiple measures, and the other a measure dimension, a conversion will also have to take place before the integration, as the two types may not be mixed in the same dataset, according to the RDF Data Cube specification.

RDF Record Table requires no choice for how to handle multiple measures, as no distinction is made between measures and dimensions. Each record simply has a number of cells, each cell with a single value. Where measured values belong together, such as in the case of a multi-spectral measurement, they can be grouped together in their own Record Table, which can be nested within the larger Record Table.

5) Ease of annotation

Setting up an RDF Data Cube requires a certain level of technical knowledge of the model. While a data entry tool can of course hide away all the complexities of the RDF itself, the user must, at the very least, specify their Dimensions and Measures. For nice, regular examples, such as those given on the RDF Data Cube website, learning how to do this is perhaps not so hard. But for more complex examples, it is asking quite a lot of the user to be able to do this correctly. It is of course possible to choose the approach of having an expert design a template for users to fill in (as in RightField [15]). The users themselves are then not required to understand the model. However, this limits the spontaneity and creativity of the users, they cannot make a simple change such as adding a new data column without needing to apply for a template change.

For RDF Record Table, on the other hand, the user does not need to make the distinction between Measures and Dimensions. All they need to do is to annotate headers with quantities, phenomena or units of measure. As the difference between a quantity and a phenomenon is not dependent on their role in the table, it is quite easy to learn. In the Rosanne tool, which implements RDF Record Table, and which we will discuss in Section VII, even this knowledge is not necessary, as the user simply looks up the annotation they want to apply, based on the name of their item.

6) Ease of integration

In RDF Data Cube, 'tables' and 'records' don't exist, the data is all merged into the hypercube. To integrate data from two different data cubes together on a given JOIN field, first the dimension to be used as the JOIN field must be chosen, optionally values of additional dimensions must be specified to select a section of the data, and finally the desired measures must be selected. For example, for the life expectancy table, we could specify Region as the JOIN field, the dimension value 2004-2006 to select that time frame, and then for the measure the only option is Life Expectancy. Given a table of average weight for the same time frame and region, we could then select the measure Average Weight, and so produce a table showing the average weight and life expectancy for all regions in the time period 2004-2006. Inherent to the integrity constraints of Data Cube is that we could not have done this if the average weight was only available for half the regions, without an extra step to generate empty 'missing data' observations for the other regions. Similarly, if the data we had available on average weight was not split into male and female, the integration could not occur, as the gender dimension would be missing in part of the integrated table. The available options are limited by the constraints placed on the data.

RDF Record Table is built around tables and records. We integrate using SPARQL [27], a semantic web querying language. To carry out the same integration as above, we select all records containing Region from the desired tables, and select the Life Expectancy and Average Weight cells. To define the time period, we set the value of that entity to 2004-2006 (remember, we have assumed that these time periods are identified entities). Missing data can be accounted for using the SPARQL OPTIONAL keyword (see Section VIII), and we can still integrate average weight information even if it is not split into male and female (the weight information is organized per region and time period, with additional nested tables containing the life expectancy per gender group). In addition to this, if desired we could join tables based on a numerical value, such as the value of the life expectancy, instead of an identifier, such as region. The distinction between Phenomenon and Quantity can guide in the selection of a join variable, but it is not required that the join variable be a Phenomenon. A table may consist solely of numerical values, if desired, and for scientific analysis such as finding correlations between variables, such a table is perfectly reasonable. RDF Data Cubes, on the other hand, must contain a dimension, and if the measured

value is turned into a dimension, then it may only take predefined values. There is much more freedom in how to integrate the data when using RDF Record Table.

7) Ease of searching and viewing

RDF Data Cube includes a data structure definition. This immediately supplies information about the expected elements in the data and its structure, making search very easy. The concept of slices allows for a particular view on the data to be quickly obtained, and the concept of dimensions makes the definition and selection of a particular slice very simple. The regularity of the data also aids search, if all integrity constraints are fulfilled then the search does not need to handle missing or optional data.

RDF Record Table does not require a data structure definition. This increases flexibility, but means that the data must first be searched to discover what types of observations are available. The use of `rec:ContainsAsExplanation` to indicate headers at the top level of the table can help make this search quicker. As there is no slice concept, pre-prepared views cannot be provided, and the absence of a Dimension concept means that selection of a given 'slice' is more complex, requiring constraints in the query. As the data is not constrained to be regular, the search query must also handle missing data and nested Record Tables. This adds to the complexity and probably reduces the speed of search.

8) Flexibility of data analysis

The requirement to choose *a-priori* between dimensions and measures is useful in fields such as standard statistics, where it is very clear what data is to be gathered. Defining dimensions and measures makes it easier to gather the data, and easier to define particular views. This requirement is, however, often problematic, particularly in research, where it is often not clear in advance what is going to be measured. Rather than having a specific measure that is influenced by certain dimensions (time, place, gender), it is often the task of a scientific study to determine what the relationship is between various measurements. Depending on the purpose of the study, the same measurement may assume the role of a cause or a consequence. Rather than assuming some causal order between quantities, therefore, it is more appropriate to simply state that they have been observed together. This is particularly the case for in-vivo studies, where it is much more common to observe various variables and try to discover their relationship, than to vary one particular variable to discover its effect, as it is often impossible to set the values of certain variables to fixed points (as is the requirement for dimensions).

We conclude that RDF Record Table can be viewed as a generalized RDF Data Cube, making fewer assumptions about the regularity and completeness of the data. If a dataset that was originally drafted as an RDF Record Table meets certain requirements, it is in principle possible to automatically transform it into an RDF Data Cube. Any dataset expressed in RDF Data Cube, on the other hand, can be modeled as RDF Record Table. This has the great advantage of allowing data in the Record Table format to still take advantage of all the tools available for Data Cube, where the data meets the Data Cube requirements. It is quite conceivable that both models could be used in the course of the same study. RDF Record Table is appropriate during the research process when data can be incomplete, the researchers are still building their understanding of the data and the role of the different factors, and diverse datasets are being integrated together. RDF Record Table then gives the researchers maximum flexibility to carry out their work without worrying about constraints, dimensions and measures. RDF Data Cube is appropriate when the data has been processed and cleaned up, and the roles of dimensions and measures are clear. RDF Data Cube then allows the researchers to take advantage of the available Data Cube tools for visualization, and to define slices of their data to make consumption and publication easier.

VII. Annotation implementation

In the following sections, we discuss two tasks that benefit from the definition from a formal model of tables and RDF Record Table in particular. In this section, we discuss how annotation of two-dimensional tables can be done in practice. This annotation is a necessary precursor for the transformation of the two-dimensional table to the RDF Record Table model. In the next section, we discuss practical support for the data integration task.

A good model of tabular data is useless if the data cannot easily be input. Given the popularity of the classic table format in tools such as spreadsheets, it should be possible to use these for data entry and then construct semantic datasets from there. In order to make this process as easy as possible, it should fit into existing work procedures and tools, and minimize additional effort by the user. Since Microsoft Excel is extremely popular, we have implemented the RDF Record Table model as an add-in for Excel, called Rosanne [25]. Rosanne supports engineers and scientists in creating semantic tables (as yet simple, non-nested, non-pivoted tables, i.e., rectangular with one header row or column, with no data values in the header). Similar functionality for the RDF Data Cube has been implemented in TabLinker [14]; however, this is a standalone tool which cannot be accessed from within Excel. Rosanne allows users to enter their data in a simple table format. Rosanne then uses OM (Ontology of units of Measure and related concepts) [14] to assist users in adding relevant quantities and units of measure to the table. In addition, other domain-specific ontologies are available for annotating identified entities in the table, such as samples, objects, locations, etc.

Support for table annotation takes two slightly different forms. In the first case, when creating and filling new, initially empty tables, the user must be assisted in selecting and assigning the right concepts and constructing the right layout. Rosanne supports this task of creating and semantically enriching tables from scratch. It does not confront the user with the Record Table model, nor does the user have to have any knowledge of ontologies. The user sets up a table by simply drawing areas in the spreadsheet. Next, the user selects the concepts they want from dropdown lists showing the user-friendly labels from the ontologies. The URIs (Uniform Resource Identifiers) for the ontology concepts are then stored in the Record Table model.

The second form of annotation is when existing datasets have to be semantically enriched. Rosanne can automatically annotate existing data with units and quantities from OM, based on heuristics [28]. This does not always produce accurate results, but saves time for the user by creating an initial annotation that can be corrected where necessary.

In addition to annotating the content of the cells in tables, a tool that handles spreadsheet tables also has to make an interpretation of the structure of a table. It needs to translate the two-dimensional form into the graph-based RDF Record Table model. Human readers can usually quickly combine layout and text in tables to make the proper interpretation. However, this is not a trivial task to automate since it depends on implicit knowledge. For the current, simple form which we support in Rosanne, an indication of the table and header areas by the user, combined with some basic assumptions in the software, suffice for the majority of tables. However, for more complex structures, more advanced processing is required. We now discuss some heuristics that could potentially be applied to make this translation.

Automatic interpretation of two dimensional tables could be facilitated by making a number of choices and assumptions on the interpretation of the table layout [29].
An important assumption, for example, is that two dimensional tables consist of rectangular blocks with cells that belong to the same semantic category, for example, they are of the same type or they can all be related to a single concept. The measured values of observations, i.e., usually numerical values of type 'float', are often grouped together in one or more blocks. In the example table on hydropower, this is the block in the lower right corner. The blocks adjacent to these float blocks, are usually of type 'string' and provide contextual information on the cells in the float blocks. In the example table, these are the two upper rows, and the column on the left side. These string blocks either represent the quantity that is measured, or the phenomenon of which that quantitative property is measured.

Another assumption is that every observation in a table can be related to a quantity and a phenomenon in the nearest string block. The string cells describing quantities can usually be recognized by the associated units of measure. Automatic recognition of quantities and units of measure can be supported by using an ontology like OM [25], and heuristics such as those used in [28], for example that units are often placed between brackets ‘Mass (kg)’. The string cells describing phenomena are usually located across from the quantity cells.

In the example table, the measure '37.76' can be related to the quantity 'Budget' and to the phenomenon 'Twenty-Second in the Lancang River Cascade'. If an observation can be associated to multiple quantities or phenomena, this could indicate that the corresponding table has a nested structure. In the example table, the measure '60' can be related to the quantity 'Installed Capacity' and to the phenomenon 'Twenty-Second in the Lancang River Cascade', but also to the phenomenon 'Realisation level', indicating nesting.

The string blocks in two dimensional tables are often called table headers, based on their position in the table. However, in RDF Record Table header cells are defined based on their role as descriptive item. Translation from a header cell in a two dimensional table to a header cell in RDF Record Table is therefore not straightforward.

Headers in two dimensional tables often contain a series of instances of phenomena or quantities. These are in fact data values (see section V) and the corresponding cells should therefore be modeled as data items in RDF Record Table. The actual header, i.e., descriptive, item in RDF Record Table is the parent class of these instances. Automatic recognition of these parent classes can be supported by using selected ontologies, for example OM for quantities and a domain vocabulary for phenomena.

The abovementioned assumptions can be used as indication of the composition of records, and properties and roles of observations when translating a two dimensional table into an RDF Record Table. However, science and engineering tables can have complex structures that are difficult to interpret in a fully automated way. A possible solution would be to develop an interactive tool. With such a tool, the majority of the interpretation would still be performed automatically, but user input is required for checking and refining the results.

VIII. INTEGRATION OF ANNOTATED DATA

Having discussed the annotation task in some detail, we now move to another important task for data handling, namely *integration* of data. Scientific research regularly requires data to be combined from different sources. This may be as simple as merging two different tables from the same experiment, or as complex as integrating multiple tables, each from a different research group at a different time. Integrating these data allows researchers to discover new relationships and to increase their knowledge.

Annotated data is easier to integrate than unannotated data. It is far easier to select the correct data through the

concepts they describe and context information, than selecting them using obscure cell coordinates and strings, which are often ambiguous and incomplete. To demonstrate the use of the RDF Record Table model for data integration, we have implemented support for this task as part of the Rosanne add-in for MS Excel.

In Rosanne, the user indicates which field is used to match records together (usually called the 'JOIN field' or sometimes the 'key') and which measurements they wish to select. Once this is done, the relevant records can be found automatically based on the annotated cells, and combined automatically using the information about the table structure.

The main challenge in integrating the annotated data is to combine the data stored in the different RDF Record Tables. SPARQL was the natural choice to perform this combination as it has the necessary functionality for searching, filtering and combining data expressed in RDFS/OWL.

In SQL, the relational equivalent to SPARQL, there is a standard functionality – the JOIN concept – which allows tables to be quickly combined. SPARQL does not have an equivalent concept, as SPARQL is based around the concept of triples, not of n-ary relations. It is standard in SPARQL to retrieve triples that share a subject, predicate or object, and in this way to combine the triples. However, integration of two records requires the integration not of two triples, but of two collections of triples. Most of these triples do not contain the common identifier, the JOIN field, on the basis of which the records are to be combined. The integration of tables is therefore more complex than the combination of isolated triples.

It was necessary for us to implement the JOIN functionality ourselves using building blocks from SPARQL, which was not a simple task. This is an important exercise for the Semantic Web, as it is becoming more and more common that tabular data is stored in RDF. We have developed a generic approach that is independent of the specific details of the tabular model, and therefore, which can work for both RDF Record Table and RDF Data Cube.

When integrating, there can be multiple records that have the same value for the JOIN field. For example, repeated measurements on the same sample. These multiple records must then be grouped together. For example, if we are joining records on the basis of the name "Jan", then the records "Jan, Wageningen, Tuesday" and "Jan, China, Tuesday" would be grouped together. To turn these records into one record, all fields except the JOIN field (which is by definition the same) must be aggregated.

Our approach follows these simple steps:

1. Select all relevant records (records containing the JOIN field)
2. Retrieve the desired information from the records
3. Group the records based on the JOIN values
4. Aggregate the other values
5. Structure all retrieved information into an integrated table
6. Retrieve the results

Steps 1 to 5 can be carried out within a single SPARQL integration query. This is a CONSTRUCT query, which creates a new RDF graph. The CONSTRUCT query nests three SELECT subqueries, which retrieve sets of variables from the existing RDF data. The innermost subquery selects the relevant records by looking for records containing an annotation that references the JOIN field (step 1). Optionally, the data to be included can be filtered in this step by using the SPARQL FILTER function, for example, we may only wish to integrate samples with a mass greater than 10g.

The second subquery selects the desired fields from the original records by looking for annotations with these fields (step 2) in the selected records, and groups the information based on the JOIN field (step 3). The outer subquery aggregates the data (step 4). Finally, the CONSTRUCT query forms the new records and creates the integrated table (step 5).

At this point, the integration is complete. However, the table is still in RDF, and is stored in the repository. To retrieve the results for recreating the two-dimensional table we use a second SPARQL SELECT query (step 6).

Note that, if wished, we could build the integrated table by simply collecting the data without aggregating them. The aggregation method, needed to construct a simplified, two-dimensional view, could then be specified when retrieving results, allowing different users to choose different views on the data. Either approach can be used depending on the situation.

It is possible that we may wish to join on more than one field. In the example above, we may not want the records about "Jan" to be merged if "Jan" is in different places. In that case, we need to identify the entities to join using both name and location.

The queries we have designed work with any number of tables. Naturally, there can be performance issues with large amounts of data.

As previously mentioned, a common challenge in scientific data is that of handling *missing data*. When collecting records from different tables, we expect to find all available records in the result, even when data (in RDF Record Table values of `rec:containsContent`) is missing. By default, however, SPARQL expects all requested information asked for in the query to be present, otherwise no result will be returned. We solved this by use of OPTIONAL clauses in SPARQL. OPTIONAL allows a section of the requested data to be missing without preventing other results from being returned. A disadvantage of OPTIONAL is that it is slower, a known performance problem of this construct [30].

A specific case of missing data is when the JOIN field mentioned in a query is missing in the data. This is more likely to occur when there are multiple join fields – in our example the name "Jan" may always be filled in, but not always the location ("Wageningen" or "China"). In this situation, the way one merges different records depends on how he or she wishes to interpret the data. One option is that the user only wants to integrate records with the same name if the location is also the same, or if the location is missing in

both. Alternatively, the user may wish to interpret the absence of location information as meaning 'any location', so that records with the same name will be integrated if both locations are the same, or if one or both locations are missing. Either choice can be catered for in the query. Our default is that records will only be matched if locations are identical. In addition, we implement the 'any location' interpretation by combining records with empty fields with all possible values of those fields, thus allowing integration with any value of that field. This is done within the integration query by means of an additional subquery.

We are currently investigating how we can support more complex queries, meeting the requirements that some users have specified for their practical cases. For example, in one situation, a scientist needs to 'integrate a measurement on a sample with the first record in time for that same sample, *after* the temperature of the sample has first exceeded 90 degrees Celsius'. For this purpose, we use the subquery facility in SPARQL to add further layers of nesting to the query. We have implemented such queries in a separate experimental tool and are now working on improving their performance and incorporating them into our main integration query. Supporting this type of queries will provide a significant benefit to researchers, as currently these integrations require a great deal of work and often specialized software or databases. How to provide a clear, intuitive user interface for such complex queries is an important issue.

All above-mentioned queries are independent of the precise tabular format. We have tested our basic integration approach developed for RDF Record Table on data in the RDF Data Cube format. The steps and the structure of the queries remain the same. The selection of the data fields is simply changed to use Data Cube syntax instead of RDF Record Table. The queries then work as designed.

For practical application of semantic integration functionality to be widely accepted, it has to be part of familiar, existing tools. Therefore we have incorporated it into our Excel add-in, Rosanne, extending the annotation functions presented before. Fig. 5 shows an example from food science. In this experiment, the researcher wishes to combine rheological measurements on protein samples with sample composition data. Without semantic support, this task would require her to find the relevant files somehow, then to copy and paste different data by hand, with plenty of scope for error. With semantically annotated tables, the necessary information is available to allow her to find the files via a search function (implemented in a demonstration tool but not yet incorporated into Rosanne). The tables have been annotated using OM and a domain ontology. The Integration Pane provides a list of all the concepts available in the files. The researcher selects 'Protein' as the identifier, and 'Storage Modulus' and 'Composition' as the variables of interest. Rosanne writes the RDF Record Table representations of the tables to a Sesame [31] repository, creates a SPARQL [27] CONSTRUCT query to find the relevant data, and generates the integrated table in the RDF Record Table format. A SPARQL SELECT query retrieves the data from the integrated table and writes it into a new

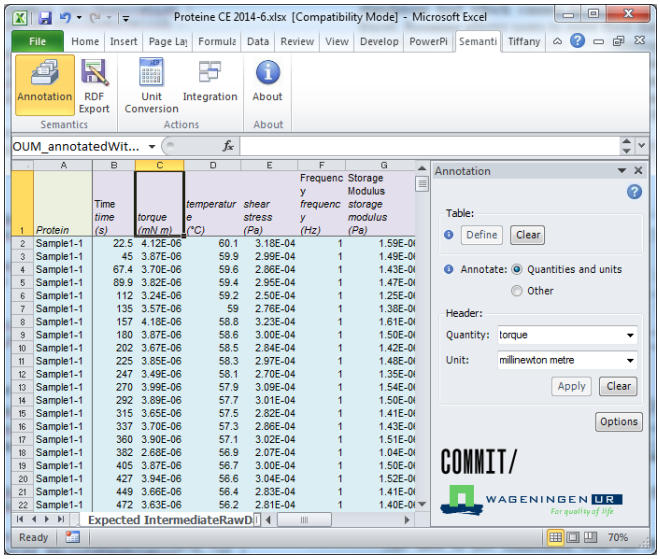


Figure 5: Rosanne using RDF Record Table.

Excel spreadsheet. The integrated table contains all the original annotations, and can itself again be integrated with other tables.

The process for the user is quite simple. She defines the integration they want with a series of simple dropdowns, and does not need to be aware of RDF Record Table, Sesame or SPARQL.

IX. Evaluation of Annotation and Integration on Industrial Use Cases

We tested annotation and integration via Rosanne on ten real-life use cases collected from four different academic research institutes, and the R&D departments of three commercial firms. These cases did not involve nested tables, but did include integration of more than two files, missing data and missing JOIN fields.

Regarding annotation, our key finding was that the provided data required some manual cleaning prior to annotation in Rosanne. Issues included JOIN variables being indirectly specified, for example, in the spreadsheet name, rather than being included in the table, tables being split over multiple locations in the spreadsheet, empty cells that were intended to be interpreted as including repetitions of previous cells, etc. Such issues can be addressed by adding data cleaning facilities to Rosanne, but are also related to compliance to good data notation by users. If information is completely missing or obscured, no tool will be able to recover it.

The integration function offered by Rosanne worked as desired for the majority of use cases once the data had been cleaned and annotated. The need for the more advanced integration queries as discussed in the previous section was confirmed by some of the other cases. These cases show how complex integration functionality, which would normally require researchers to turn to specialist solutions

such as databases, can in future be offered in Excel in a simple, user-friendly manner.

One case involved a pivoted table, a table where the header contains data elements. As Rosanne does not yet support annotating the data elements within the header of a pivoted table, this table had to be unpivoted before it could be annotated. For the integration step itself, a wider range of data aggregation would be welcome. Some use cases, however, required advanced statistical analysis methods, such as regression, which fall outside the scope of aggregation. We are looking at a possible combination of Rosanne and a statistical package like R, rather than attempting to support this sort of analysis in SPARQL queries.

We conclude from these tests that RDF Record Table and our SPARQL approach for integration were successful in carrying out integrations on real-life data. While additional functionality is necessary to achieve the full results desired in some use cases, the core integration functionality was shown to be sound for a variety of data.

Due to the manual cleaning required, it was not possible to fully validate the performance and practicality of Rosanne in these use cases. Once the issues discovered in these use cases have been addressed, we will conduct a full validation of our approach.

X. CONCLUSION AND FUTURE WORK

We have proposed RDF Record Table as a way to model heterogeneous tabular data semantically. The model complements the RDF Data Cube vocabulary. RDF Data Cube offers the benefits of semantic modelling to domains such as statistics, with regular, standardized datasets, and provides good support for data visualization. RDF Record Table offers more flexibility in storing heterogeneous or incomplete data, and therefore extends the benefits of semantic modelling to the more complex situations for which RDF Data Cube is too restrictive. RDF Record Table addresses all four aspects identified in Section I as being essential for a good tabular model; the content can be annotated, the table structure is modelled, there is a link to the PROV model for provenance data, and it is flexible, allowing complex structures.

A first implementation of the RDF Record Table model as an extension of Microsoft Excel, called Rosanne, demonstrates that the format is capable of accurately representing tabular data, and can be applied by offering users simple choices from drop-downs, without the users needing to be aware of the RDF Record Table itself.

Rosanne also provides semi-automatic integration of datasets. SPARQL queries are used to integrate data from different RDF Record Tables. This integration approach is defined in a generic way, making it applicable to other RDF tabular models, such as RDF Data Cube. The user can specify their integration using simple drop-downs, and again does not need to be aware of the complexity of the model or the queries. This integration functionality has been evaluated in use cases from a number of research institutes and R&D organizations of multinationals in food production, cooperating in TI Food and Nutrition [32]. While a number of issues were identified that must be addressed to make Rosanne a practical tool for industry, the core integration principles were shown to be sound.

In Section I, we discussed the various problems that arise from how spreadsheet data is currently handled. RDF Record Table and its implementation in Excel provide a means to effectively tackle these problems. Ambiguity and incomprehensibility are addressed by linking data to defined concepts in shared vocabularies. The link between RDF Record Table and the PROV model allows the provenance of the data to be recorded. The annotations can be searched, making it easier to locate relevant data. The integration facility of Rosanne, built on top of the RDF Record Table model, enables data from different spreadsheets to be linked and combined together, assisting reuse. Finally, this support is available in the commonly used Excel tool, allowing researchers to incorporate good data bookkeeping into their research workflow, thus enabling good data documentation with minimal effort.

For full implementation of the RDF Record Table model, several issues must still be solved. While the model itself supports nesting, the Rosanne add-in does not. This support must be added, preferably by offering heuristics that assist the user. To handle the overhead of explicit annotations, RDF Record Table allows repeated information to be presented in cells that provide metadata for similar cells. However, methods for automatic (local) expansion and compression of datasets should be considered as well. In order to realize the benefits of both RDF Record Table and RDF Data Cube on the same data at different stages in the scientific or engineering process, mapping between the two formats is required. This necessitates support for implementing the constraints of Data Cube when converting Record Table to Data Cube.

As discussed, the integration functionality of the Rosanne add-in can be improved by allowing more complex queries, handling nested tables and offering a link to a statistical package for advanced data analysis.

Rosanne has not yet been optimized for performance on large datasets. This optimization will be a necessary step in producing an add-in that can be used in industry.

In addition to these issues on the annotation and processing of new data, the recovery of legacy data needs attention. There is a wealth of data stored in existing spreadsheets, which have, in general, an informal structure and no annotations. Current results for fully automatic annotation are still of insufficient quality [28], so more research is needed to find how to unlock this legacy data.

Semantic tables also offer the potential to support cleaning of the data, for example by defining allowed units and ranges for measurements so that errors can be detected and possibly (semi-)automatically corrected. This is an aspect that we will look at in the future.

A format for tabular data is of little use if it is not adopted by the community. We plan to submit RDF Record Table to the CSV on the Web Working Group [33] for consideration and inspiration in their work to provide better interoperability for tabular data.

Describing the content and structure of tabular data semantically makes it possible to easily find data even in disparate sources, to understand and clean the data and to combine it semi-automatically. This way, much richer datasets will be published in the future, so that others can fully understand them and build further on them.

ACKNOWLEDGMENT

This publication was supported by the Dutch national program COMMIT.

REFERENCES

[1] J. Top, H. Rijgersberg, and M. Wigham, "Semantically enriched spreadsheet tables in science and engineering," in Proc. Eighth International Conference on Advances in Semantic Processing (SEMAPRO), pp. 17-23, Rome, Italy, 2014.

[2] Y. L. Simmhan, B. Plale, and D. Gannon. "A survey of data provenance in e-science," ACM SIGMOD Record, 2005. doi:10.1145/1084805.1084812.

[3] A. Garcia, O. Giraldo, and J. Garcia, "Annotating Experimental Records Using Ontologies," Int. Conference on Biomedical Ontology, Buffalo, NY, USA, 2011. Available from: http://ceur-ws.org/Vol-833/paper12.pdf. Retrieved June, 2014.

[4] J. Gray, "Jim Gray on eScience: a transformed scientific method," in T. Hey, S. Tansley, K. Tolle (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, 2009, pp. xvii–xxxi.

[5] RDF Semantic Web Standards, W3C. Available from: http://www.w3.org/RDF/. Retrieved: May, 2015.

[6] R. Cyganiak, D. Reynolds, (eds). RDF Data Cube Vocabulary, W3C, 2012. Available from: http://www.w3.org/TR/vocab-data-cube/. Retrieved June, 2014.

[7] L. Han, T. Finin, C. Parr, J. Sachs, and A. Joshi, "RDF123: From spreadsheets to RDF," Lecture Notes in Computer Science, Vol. 5318 LNCS, 2008, pp. 451–466. doi:10.1007/978-3-540-88564-1-29.

[8] J. Cunha, J. Saraiva, and J. Visser, "From spreadsheets to relational databases and back," In Proceedings of the 2009 ACM SIGPLAN workshop on Partial evaluation and program manipulation - PEPM "09 (p. 179), 2009.

[9] C. Bizer, and R. Cyganiak, "D2R Server – Publishing Relational Databases on the Semantic Web," World, p. 26, 2006.

[10] L. Lefort, J. Bobruk, A. Haller, K. Taylor, and A. Woolf, "A Linked Sensor Data Cube for a 100 year homogenised daily temperature dataset," In Proceedings of the 2012 5th International workshop on Semantic Sensor Networks – SSN 2012 (p. 1), 2012.

[11] S. Capadisli, S. Auer and A.-C. Ngonga Ngomo, "Linked SDMX Data," Semantic Web, 2013. doi:10.3233/SW-130123.

[12] Commercially Empowered Linked Open Data Ecosystems in Research project. Available from: http://code-research.eu/. Retrieved March, 2015.

[13] Tabels Project, Available from: http://idi.fundacionctic.org/tabels/. Retrieved March, 2015.

[14] TabLinker, 2012. Available from: http://www.data2semantics.org/2012/02/19/tablinker/. Retrieved June, 2014.

[15] RightField. Available from: https://www.sysmo-db.org/rightfield. Retrieved March, 2015.

[16] P. Rivera Salas, F. Maia Da Mota, M. Martin, S. Auer, K. Breitman, and M. A. Casanova, "Publishing Statistical Data on the Web," 2012 IEEE Sixth International Conference on Semantic Computing, Palermo, 2012. doi: 10.1109/ICSC.2012.16.

[17] OpenCube Toolkit. Available from: http://opencube-toolkit.eu/. Retrieved March, 2015.

[18] OpenCube Project. Available from: http://opencube-project.eu/ . Retrieved March, 2015.

[19] L. Lefort, H. Leroux, "Design and generation of Linked Clinical Data Cubes," First Internatianal Workshop on Semantic Statististics, Sydney, Australia, 2013.

[20] P. Groth, L. Moreau. (eds), PROV Overview, W3C, 2013. Available from: http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/. Retrieved June, 2014.

[21] M. Nilsson, A. Powell, P. Johnston, and A. Naeve. "Expressing Dublin Core metadata using the Resource Description Framework (RDF)," 2008. Available from: http://dublincore.org/documents/dc-rdf . Retrieved June, 2014.

[22] Matlab, The Language of Technical Computing. Available from: http://www.mathworks.nl/products/matlab/. Retrieved July, 2014.

[23] SPSS Statistics. Available from: http://en.wikipedia.org/wiki/SPSS. Retrieved July, 2014.

[24] The R Project for Statistical Computing. Available from: http://www.r-project.org/. Retrieved March, 2015.

[25] H. Rijgersberg, M. Wigham, and J. L. Top, "How semantics can improve engineering processes: A case of units of measure and quantities," Advanced Engineering Informatics, 25(2), 2010, pp. 276–287. doi:http://dx.doi.org/10.1016/j.aei.2010.07.008.

[26] R. Hodgson, P. J. Keller, J. Hodges, and J. Spivak, "QUDT - Quantities, Units, Dimensions and Data Types Ontologies," Available from: http://qudt.org/. Retrieved June, 2014.

[27] W3C, SPARQL Query Language for RDF. Available from: http://www.w3.org/TR/rdf-sparql-query/. Retrieved July, 2014.

[28] M. van Assem, H. Rijgersberg, M. Wigham, and J.L Top, "Converting and annotating quantitative data tables," The Semantic Web - ISWC 2010, vol. 6496/2010, 2010, pp. 16–31. doi:10.1007/978-3-642-17746-0_2.

[29] M.G. De Vos, W. R Van Hage, J. Ros, A.T. Schreiber, 2012. "Reconstructing Semantics of Scientific Models : a Case Study," In Proceedings of the OEDW workshop on Ontology engineering in a data driven world, EKAW 2012. Galway, Ireland.

[30] A. Loizou, R. Angles, and P. Groth. "On the Formulation of Performant SPARQL Queries," Web Semantics: Science, Services and Agents on the World Wide Web. Vol. 29, 2014 doi:10.1016/j.websem.2014.11.003.

[31] Sesame framework for RDF data. Available from: http://rdf4j.org/. Retrieved March, 2015.

[32] TI Food and Nutrition. Available from: http://www.tifn.nl. Retrieved July, 2014.

[33] CSV on the Web Working Group Charter, 2013. Available from: http://www.w3.org/2013/05/lcsv-charter.html. Retrieved June, 2014.

A Decision Support Approach based on Sentiment Analysis Combined with Data Mining for Customer Satisfaction Research

Nafissa Yussupova, Maxim Boyko, and Diana Bogdanova
Faculty of informatics and robotics
Ufa State Aviation Technical University
Ufa, Russian Federation
Emails: {yussupova@ugatu.ac.ru, maxim.boyko87@gmail.com, dianochka7bog@mail.ru}

Andreas Hilbert
Faculty of Economics
Dresden University of Technology
Dresden, Germany
andreas.hilbert@tu-dresden.de

***Abstract*—This paper describes the application of a novel domain-independent decision support approach for Customer Satisfaction Research. It is based on customer satisfaction research through deep analysis of consumer reviews posted on the Internet in natural language. Artificial Intelligence techniques, such as web data extraction, sentiment analysis, aspect extraction, aspect-based sentiment analysis and data mining, are used for realization of consumer reviews analysis. In paper, specific Internet resources (such as yelp.com, tripadvisor.com, tophotels.ru) are used for accumulating customer reviews as a data source. This is performed in accordance with the quality standard ISO 10004 and proposed decision support approach, which allows for both qualitative and quantitative customer satisfaction surveys to be carried out. The output of the quantitative survey are values of customer satisfaction with product and each product's aspect. The output of the qualitative survey are significance values of products aspect for customers and identified latent relations between overall satisfaction with product and satisfaction with products' aspects. The proposed approach is performed as a prototype of a decision support system. To evaluate the efficacy of the proposed approach, two experiments on hotels and banks customer reviews have been carried out. The obtained results prove the efficacy of the proposed decision support approach for quality management and the concept of using it instead of classical methods of qualitative and quantitative research of customer satisfaction.**

Keywords-customer satisfaction research; decision support system; sentiment analysis; data mining.

I. INTRODUCTION

In order to provide product quality, a company should make effective managerial decisions. In the modern world, the efficacy of managerial decision-making process depends on the information available to the person that makes decisions and the depth of information analysis. Therefore, a company should develop processes of automated collection of information and its further processing and analysis. Decision-making should be based on the knowledge and principles obtained during the analysis of the collected data. In this article, we expand on our research work presented at The Third International Conference on Data Analytics (2014) [1].

Quality assurance is currently attained through a process approach based on the model of a quality management system [2] (see Figure 1). It describes the interaction of the company and the customer during the process of product production and consumption. To correct the parameters of a product's quality in order to improve it for the customer, the model has feedback. For companies, feedback during the process of quality management is the information about the level of customer satisfaction, which is expressed in the form of customer reviews about a product's quality. That is why customer satisfaction is key information for quality management that influences decision-making.

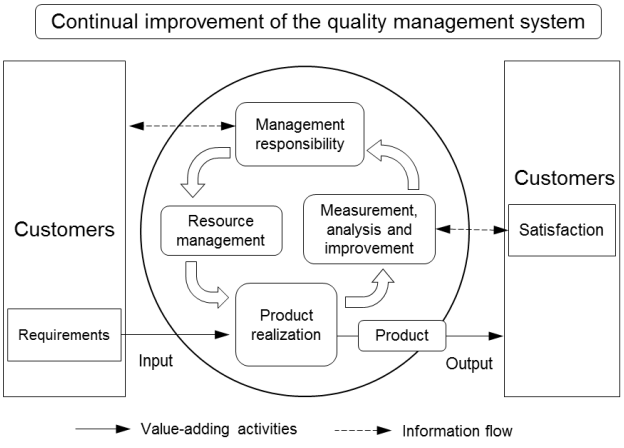


Figure 1. Model of a process based quality management system.

To collect data and evaluate customer satisfaction, International Quality Standards ISO 10004 (International Organization for Standardization) recommends using the

following classical methods: face to face interviews, telephone interviews, discussion groups, mail surveys (postal questionnaires), on-line research and surveys (questionnaire surveys) [3]. However, these methods of collection and analysis of customer opinions have a number of significant drawbacks.

A general drawback of these methods is a large amount of manual work: preparing questions, creating a respondent database, mailing questionnaires and collecting results, conducting a personal interview, and preparing a report. All of these procedures make a research expensive. These methods cannot monitor customer satisfaction continuously. For this reason, monitoring is limited by a one-time period, because costs rapidly grow with an increase in the frequency of monitoring. There is no possibility for monitoring trends of customer satisfaction. It also has a negative influence on lengthiness of managerial decision making.

Another problem regards various scales for measuring customer satisfaction and their subjectivity perception. Value of customer satisfaction is estimated by abstract satisfaction indices that are difficult to understand, hard to compare and interpret. Furthermore, methods for data analysis recommended by ISO 10004 [3] allow detection of only linear dependencies and relations in data, such as linear correlation. Using only linear methods researcher could miss the impact of the mutual satisfaction of the product's aspects to the overall satisfaction with a product. In paper, an aspect means characteristics, attributes, and properties that characterize the products, e.g., a "phone battery" or "delivery period".

The aim of this paper is the development of a decision support approach for quality management provided by analysis of customer reviews on the Internet with use of web data extraction, sentiment analysis, aspect extraction, aspect-based sentiment analysis and data mining that could overcome the aforementioned drawbacks of the classical approach for customer satisfaction research. The main contribution of our research is novel approach using results of sentiment analysis for further knowledge extraction about dependencies between overall satisfaction with product and product's aspects.

The remainder of this paper is organized as follows: in Section II, we focused on overview of recent solutions and frameworks for analysis of user generated content and their drawbacks. In Section III, we described architecture and workflow of proposed decision support system. In Section IV, we described using text mining and data mining techniques for qualitative and quantitative customer satisfaction surveys. In Section V, we provide two experiments of customer satisfaction research: 1) qualitative and quantitative surveys for two hotels and whole resort and 2) qualitative survey for Russian banks.

II. Related Work

Applying Text Mining tools for analyzing customers' reviews posted on the Internet is not novel. There are many studies concerning models and methods for data collection, sentiment analysis and information extraction. Recent studies show acceptable accuracy of methods for sentiment classification. Gräbner et al. [4] proposed a system that performs the sentiment classification of customer reviews on hotels. The precision values are 84% for positive and 92% for negative reviews. Lexicon-based method [5] allowed the correct classification of reviews with a probability of about 90%. These achievements make sentiment analysis applicable for an application on quality management and customer satisfaction research.

Jo and Oh [6] and Lu et al. [7] considered the problems of automatically discovering products' aspects and sentiments estimation for these aspects, which are evaluated in reviews. For solving these problems, they suggested methods based on Latent Dirichlet Allocation [8] and its modifications.

A lot of social monitoring systems and frameworks have been developed for automatic analysis of reviews and topics. Liu et al. [9] presented framework called Opinion Observer for analyzing and comparing consumer opinions of competing products. This prototype system is able to visualize the strengths and weaknesses of each product in terms of various product features. For visualization it use actual number of positive or negative opinions normalized with the maximal number of opinions on any feature of any product. For experiments, authors used reviews on electronic products. Kasper and Vela [10] presented a web based opinion mining system for hotel reviews and user comments that supports the hotel management called BESAHOT. The system is capable of detecting and retrieving reviews on the web, classifying and analyzing them, as well as generating comprehensive overviews of these comments. Ganu et al. [11] focused on an analysis of free-text reviews by means of classification of reviews at the sentence level, with respect to both the topic and the sentiment expressed in the sentences. For experiments, authors used reviews on restaurants. Blair-Goldensohn et al. [12] proposed a system that summarizes the sentiment of reviews for a local service, such as a restaurant or hotel. In particular, they focus on aspect-based summarization models, where a summary is built by extracting relevant aspects of a service, such as service or value, aggregating the sentiment per aspect, and selecting aspect-relevant text. Bjørkelund et al. [13] described how the results of sentiment analysis of textual reviews can be visualized using Google Maps, providing possibilities for users to easily detect good hotels and good areas to stay in. Ajmera et al. [14] developed a Social Customer Relationship Management (SCRM) system that mines conversations on social platforms to identify and prioritize those posts and messages that are relevant to enterprises. The system aims to empower an agent or a representative in an enterprise to monitor, track and respond to customer communication while also encouraging community participation. Bank [15] proposed interactive Social Media monitoring system to extract related information from user generated content. One of the important contribution of this work was the proposition of new quality indeces. One of the important contribution of work was proposition of new quality indeces. The Relevancy Index states the importance of a

given topic and provides a robust marketing and market penetration independent importance information. The Market Satisfaction Index provides the possibility to compare several product features among different products or manufacturers. The Product Satisfaction Index extracts the advantages and disadvantages of a product.

In some related work, authors pay attention to relations between overall ratings of products, and ratings of products' aspects evaluated in the review. Wang et al. [16] formulated a novel text mining problem called Latent Rating Analysis (LARA). LARA aims at analyzing opinions expressed in each review at the level of topical aspects to discover each individual reviewer's latent rating on each aspect as well as the relative importance weight on different aspects when forming the overall judgment. For solving this problem probabilistic rating regression model is used. For experiments, authors used reviews on hotels. De Albornoz et al. [17] aimed to predict the overall rating of a product review based on the user opinion about the different product features that are evaluated in the review. For experiments, authors used reviews on hotels.

Wachsmuth et al. [18] formulated and validated an important hypothesis that the global sentiment score of a hotel review correlates with the ratio of positive and negative opinions in the review's text and that the global sentiment score of a hotel review correlates with the polarity of opinions on certain product features in the review's text.

The main drawback of these considered systems is that they can provide entirely only a quantitative survey of customer reviews, i.e., they can provide measurement of the degree of customer satisfaction with a product and its aspects. Qualitative survey were usually only conducting the extraction of products' aspects. However, estimation of the significance of each products' aspects for the customer is missed. The information about products' aspects that influence customers' satisfaction and relative importance of products' aspects for the customers is missing, as well as an insight into customer expectations and perceptions.

The most related work to this problem is [19]. It is dedicated to the topic of aspect ranking, which aims to automatically identify important aspects of product from online consumer reviews. Most proposals used a probabilistic model with a large number of parameters that lead to low robustness of the model. Total weighting values of aspects are calculated as the average of the weighting values by each review. Finally, significance values of aspects are estimated independently of an opinion's sentiment, e.g., in real life, we can discuss in review about bad "signal connection", but we usually omit comments in case of good "signal connection", because it must be in phone. In our paper, we estimate significance values of aspects in accordance with their positive and negative sentiments. In this manner, it is possible to use the Kano's model of customer satisfaction [20], which classifies customer preferences into four categories.

In this paper, for qualitative survey is used a novel approach based on transformation results of sentiment analysis and aspect-based sentiment analysis, such as sentiment labels of reviews and mentions about product's aspects in reviews, into boolean data. After that, boolean data is processed with a data mining tool – decision tree (see Section IV). Qualitative survey aims to identify how the sentiment of reviews depends on the sentiment of different products' aspects. In other words, how overall customer satisfaction with product depends on the customer satisfaction with a product's aspects. Decision tree performs this aim and identifies latent relations between the sentiment of reviews and sentiment of a product's aspects. Also using the decision tree allows to estimate the significance of product's aspects for the customers. Output of qualitative survey are significance values of product's aspects for customers and identified latent relations between satisfaction with product and satisfaction with each product's aspect, which produced as rules extracted by the decision tree. The availability of both quantitative and qualitative surveys allows realizing Intelligent Decision Support System for Quality Management in accordance with quality standard ISO 10004.

III. The Proposed Decision Support Approach

The suggested approach to decision making in product quality management accomplished through unification of methods for collecting and processing text data into Intelligent Decision Support System (IDSS). The architecture (subsystems and contained modules) of the obtained IDSS is presented in Figure 2. The subsystem of monitoring and data collection fills the warehouse with customer reviews and other relevant information. It also supports the actuality of data via automated monitoring of Internet resources and carries out data cleansing. In the subsystem of monitoring and data collection is realized methods of web data extraction. The data storage subsystem provides safe-keeping and integrity of collected reviews and results of data processing. In the subsystem of data analysis are realized methods of aspect extraction, sentiment analysis of reviews, aspect-based sentiment analysis, and decision tree. In subsystem of user interaction is visualized results of analysis.

In Figure 3, the algorithm of the IDSS is presented. It consists of four stages. The first stage includes collection of reviews from Internet resources, data cleansing and loading reviews into the database. IDSS is able to actualize data everyday and to correct current customer satisfaction that allows provide continuous monitoring. The second stage performs processing collected reviews. It includes preprocessing procedures, such as preparing training samples of reviews for sentiment classifier, text lemmatization, and encoding text of reviews in vector form. Processing procedures include extraction of a product's aspects, training of the classifier and sentiment analysis of reviews and aspect-based sentiment analysis.

The third stage is the quantitative survey. The quantitative survey is based on sentiment analysis of reviews entirely, and aspect-based sentiment analysis of sentences with mentions of a product's aspects. Sentiment classification is attained through binary scale – positive

and negative sentiments. As a measure of the customer satisfaction with product is used a ratio of positive reviews to the sum of positive and negative reviews. As a measure of the customer satisfaction with product's aspects is used a ratio of positive sentences with mentions of a product's aspect to the sum of positive and negative sentences with mentions of a product's aspect. The output of the quantitative survey is values of customer satisfaction with a product and each product's aspect.

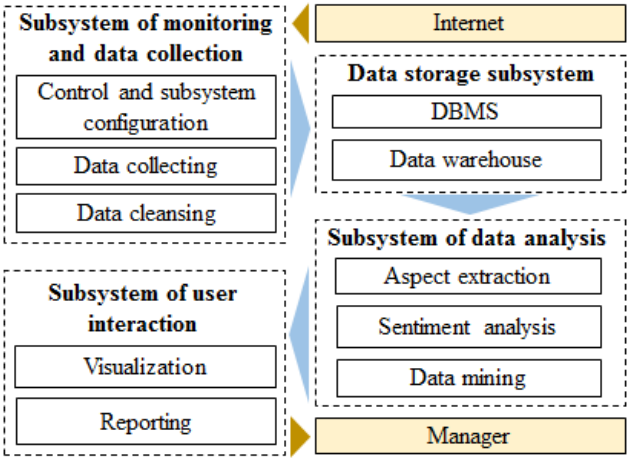


Figure 2. The architecture of Intelligent Decision Support System for product quality management.

The fourth stage is the qualitative survey of customer satisfaction. It is based on transformation results of sentiment analysis into Boolean data and following constructing of decision tree on it. The qualitative survey aims to identify how sentiment of review depends on the sentiment of different aspects of a product. Decision tree performs this aim and identifies latent relations between sentiment of a review and sentiment of a product's aspects. The output of the qualitative analysis is significance values of a product's aspects for customers and identifying latent relations extracted by decision tree. Managerial decision development and making is carried out on the basis of the performed quantitative and qualitative surveys.

IV. Applied Artificial Intelligence Techniques

In this section are described implemented AI techniques for customer satisfaction surveys and support decision making.

A. Data collection

Nowadays there are a large number of Internet resources where users can leave their opinions about products and services. The most popular examples of review sites are tophotels.ru (635 thousand reviews), yelp.com (53 million reviews), tripadvisor.com (130 million reviews). Similar resources continue to gain popularity. As opposed to social networking services, the advantage of review sites lies in their purpose - accumulation of customer reviews. One more advantage is that many of such resources have moderators of reviews and confirmation of author's objectivity, e.g., registration procedure.

1. Stage of collecting reviews
a) Collecting reviews on the Internet from sources.
b) Cleansing reviews.
c) Loading reviews into the database.

2. Stage of processing reviews
a) Forming of the training sample for sentiment classifier.
b) Lemmatization of reviews.
c) Encoding text of reviews in vector form.
d) Extraction of product's aspects.
e) Training of the sentiment classifier.
f) Sentiment analysis of reviews.
g) Aspect-based sentiment analysis of sentences with mentions about product's aspects.

3. Stage of quantitative survey
a) Estimation and comparison of customer satisfaction for own product and competitor's product.
b) Dynamic trend analysis of customer satisfaction with products and with product's aspects.
c) Identification of negative trends in customer satisfaction.
d) Defining problems in quality of product.

4. Stage of qualitative survey
a) Converting results of sentiment analysis into boolean data.
b) Constructing decision trees on boolean data.
c) Estimation of significance of the product's aspects.
d) Identification dependencies between overall customer satisfaction with product from customer satisfaction with each product aspects.

Figure 3. Working algorithm of Intelligent Decision Support System

There are two main types of collecting data from the Internet resources of customer reviews: 1) by using API (Application Programming Interface), and 2) by web data extraction. API is a set of ready-to-use tools - classes, procedures, and functions - provided by the application (Internet resource) for using in an external software product. Unfortunately, only a few resources that accumulate reviews have API.

In this paper is used the second method for data collection – web data extraction. It is a process of automated content collection from HTML-pages of any Internet resource using special programs or script. Related work is presented in [21][22]. Scheme of reviews collection is presented in Figure 4. Web pages of review sites have specific HTML-structure that includes separate blocks with the name of a product or company, author's review, and other blocks with additional information, such as date, place. Therefore, all reviews are clearly identified in relation to the object of review. It significantly simplifies the process of data collection in contrast to collecting messages from social networking services.

Input: review site

1. Gathering links on pages with reviews or generating links by using a template.
2. Setting boundaries for a content by using a HTML-structure.
3. Collection data on the set of links.
4. Cleaning data and removing duplicates.
5. Loading data into the database.

Output: set of customers' reviews *D*

Figure 4. Algorithm of web data extraction

B. Sentiment Analysis

After data collection, it is possible to process review data with Text Mining tools. In this paper automatic sentiment analysis of reviews is used to evaluate customers' satisfaction with product and product's aspects. Sentiment stands for the emotional evaluation of author's opinion about a product that is referred to in the reviews.

There are three main approaches to sentiment analysis: 1) linguistic, 2) statistical, and 3) combined. The linguistic approach is based on using rules and vocabularies of emotionality words [23][24]. This approach is quite time-consuming due to the need of compiling vocabularies, patterns, and making rules for identifying sentiments. However, the main drawback of this approach is the impossibility to get a quantitative evaluation of the sentiment. The statistical approach is based on the methods of supervised and non-supervised machine learning (ML) [25][26]. The combined approach presupposes a combined use of the first two approaches.

In this paper we used methods of supervised machine learning – naïve Bayesian classifier and Support Vector Machines. Text sentiment evaluation can be expressed quantitatively. Their realization in IDSS is based on techniques described by Pang and Lee [25][26]. More detailed information about implemented methods of sentiment analysis used in this paper can be found in our previous work [27][28]. In Figure 5 algorithms of learning and classification for naïve Bayes classifier based on Multinomial model are presented. An advantage of these ML methods that they are quite easy in software implementation, and do not require making linguistic analyzers or sentiment vocabularies. They are able to evaluate sentiment quantitatively. For sentiment classification is used binary scale - positive and negative tonality. We use vector representation of review texts with help of the bag-of-words model. As attributes, we consider bit vectors - presence or absence of the word in the review text, and frequency vectors – a number of times that a given word appears in the text of the review. Lemmatization is also used. We also used lemmatization that transforms all the words of the review to the initial form.

Learning of naive Bayes classifier
Input: training set of reviews $D' = \{(d_1, c_1), \ldots, (d_m, c_m)\}$, set of classes $C = \{positive, negative\}$.

1. Extract all words from D' to the vocabulary V
2. For each $c \in C$ do
   3. Count documents N^c in each class c
   4. Calculate probability $p(c) = N^c / N$
   5. For each $w_t \in V$ do
      6. Count number of occurrences $K^c_{w_t}$ of word w_t in each class
      7. Calculate prob. $p(w_t \mid c) = (K^c_t + 1) / \sum_{t=1}^{|V|} (K^c_t + 1)$

Output: V, $p(c)$, $p(w_t \mid c)$

Classification with naive Bayes classifier
Input: review d from set D, V, $p(c)$, $p(w_t \mid c)$

1. Extract all words from d to the vocabulary V_d
2. For each $c \in C$ do
   3. Calculate $score[c] = \ln p(c)$
   4. For each $w_t \in V_d$ calculate $score[c] += \ln p(w_t \mid c)$
5. If $score[pos] > score[neg]$ than $d \in positive$ else $d \in negative$

Output: sentimental label *Sent* of review *d* (positive/negative)

Figure 5. Algorithm of naïve Bayes classifier

C. Aspect-based Sentiment Analysis

Sentiment Analysis of reviews allows the evaluation of overall customer's satisfaction with product. However, it does not clearly show what customers like about a product and what they do not like. To answer this question, it is necessary to perform an aspect-based sentiment analysis. An aspect means characteristics, attributes, and properties that characterize the products, e.g., a "phone battery" or "delivery period". However, one product can have a great number of aspects. Furthermore, aspects in the text can be expressed by words-synonyms, e.g., "battery" and "accumulator". In this case, it makes sense to combine aspects into aspect groups.

Aspect-based sentiment analysis of the review is a more difficult task and consists of two stages – identifying all product's aspects and determining the customers' sentiment of the comment on them. To complete the task of the aspect-based sentiment analysis, we developed a simple algorithm (see Figure 6). Aspects extraction based on the frequency of nouns and noun phrases mentioned in reviews based on work [29].

A frequency vocabulary [30] (created on text corpus) that helps to compare the obtained frequencies from reviews with frequencies from corpus is used to identify aspects. The nouns with maximum frequency deviations

are claimants to be included into aspect groups. Clustering of the nouns into aspect groups was carried out manual. It should be noted, that if a sentence includes nouns from several aspect groups, then it would refer to opinion about each aspect group of these nouns.

Aspect extraction
Input: set of reviews *D*

1. Extract all nouns *S* from the set of reviews *D*.
2. Count the frequency of nouns $\forall t = 1,|S| : f_t = N_t / N$ in the whole set of reviews *D*, where *N* – number of appearances of all words, N_t – number of appearances of the *t* noun.
3. Count the difference $\forall t : \Delta_t = f_t - f_t^v$ between the counted frequencies f_i and vocabulary frequencies f_t^V.
4. Sort the set of nouns *S* in descending order Δ_t.
5. Divide the set of nouns *S* from $\Delta_t > 0$ into aspect groups.

Output: set of aspect groups and aspect words

Aspect-based sentiment classification
Input: sentiment classifier, set of aspect groups and aspect words

1. Divide a set of reviews into set of sentences.
2. Perform sentiment classification for each sentence.
3. Check each sentence for the condition: if a sentence has a sentiment score (negative or positive) greater than a threshold *h* and contains at least one noun from any aspect group, then this sentence is labeled as an opinion (negative or positive) about the given product's aspect.

Output: labeled sentences with mentions about product's aspects $\{Neg_{i1},..., Neg_{im}, Pos_{i1},..., Pos_{im}\}$

Figure 6. Algorithm of aspect-based sentiment analysis

The results of sentiment analysis and aspect-based sentiment analysis can be presented in the form of text variables $Obj = (\mathrm{Re}v_i, Sent_i, Neg_{i1},..., Neg_{im}, Pos_{1i},..., Pos_{im})$, where *Obj* – a object or a product, Rev_i – text of the *i* review, $Sent_i$ – sentiment label of *i* review, Neg_{ij} – negative sentences with mention about the *j* aspect group in the *i* review, Pos_{ij} – positive sentences with mention about the *j* aspect group in the *i* review, *i* – number of review, *j* – number of aspect group, *m* – amount of aspect groups.

D. Data Mining

The present paragraph suggests an algorithm of the following processing of results of sentiment analysis and aspect-based sentiment analysis. The aim of the developed algorithm is to discover latent knowledge that can be used for decision support in product quality management. To realize this algorithm we use the Data Mining method – decision tree, since it is easy to understand and interpret results. It also can explain relations between overall sentiment of review and sentiment of each aspect group by means of Boolean logic.

Input: positive and negative sentences with mentions about product's aspects $\{Neg_{i1},..., Neg_{im}, Pos_{i1},..., Pos_{im}\}$, vector of sentimental labels *Sent* of reviews.

1. Convert a text data $Obj = (\mathrm{Re}v_i, Sent_i, Neg_{i1},..., Neg_{im}, Pos_{1i},..., Pos_{im})$ into a boolean type by the following rules:
   2. If $Sent_i = negative$, then $newSent_i = 1$, else $newSent_i = 0$
   3. If $Neg_{ij} \neq null$, then $newNeg_{ij} = 1$, else $newNeg_{ij} = 0$
   4. If $Pos_{ij} \neq null$, then $newPos_{ij} = 1$, else $newPos_{ij} = 0$
5. Creating a decision tree where the variable $newSent_i$ is a dependent variable from $\{newNeg_{i1},..., newNeg_{im}, newPos_{i1},..., newPos_{im}\}$
6. Estimation significances of aspect groups and interpretation of extracted rules

Output: significance values of product's aspects, latent relations between satisfaction with product and satisfaction with aspects

Figure 7. Algorithm of data mining

The developed algorithm of knowledge discovery includes procedures presented in Figure 7. The main idea of this algorithm is to convert results of aspect-based sentiment analysis to Boolean data considering both positive and negative mentions about product's aspects. Then we apply decision tree for obtained Boolean data. The described algorithm allows understanding of which sentiment sentences about a product's aspects influence the overall sentiment of review or, in other words, what product aspects influence customer satisfaction and in what way. The constructed decision tree model allows the consideration of the influence on overall satisfaction with product of not only each satisfaction with some product's aspect, but also mutual presence of satisfaction and dissatisfaction with different product's aspects in the review. In other words, this approach is able to identify non-linear dependencies between overall satisfaction with product and satisfaction with product's aspects. The decision tree model also allows the detection of the most significant product's aspects that are essential for the customer.

In Figure 8 an example of decision tree model is presented. Nodes of the decision tree are the sentiment aspects' variables, i.e., presence or absence in the review sentimental sentences (positive or negative) with mention about some aspect from aspect group. Edges of the tree are the values of aspect variables, i.e., 1 is presence, 0 is absence. Leaves present overall sentiment of review, i.e., each branch leads to either a positive review or a negative review that meets customer satisfaction or dissatisfaction in dependence, which product's aspects satisfy or dissatisfy the customer.

The decision tree model can be expressed both in the form of Boolean functions (see Eq. (1)) in a disjunctive normal form, and in natural language as rules:

$$\begin{aligned} &Rule\ \#1: \ \overline{Neg.a.g.\#2} \rightarrow Pos.\ review \\ &Rule\ \#2: \ Neg.a.g.\#2 \cap \overline{Pos.a.g.\#3} \rightarrow Neg.\ review \\ &Rule\ \#3: \ Neg.a.g.\#2 \cap Pos.a.g.\#3 \cap \overline{Pos.a.g.\#1} \rightarrow Neg.\ review \\ &Rule\ \#4: \ Neg.a.g.\#2 \cap Pos.a.g.\#3 \cap Pos.a.g.\#1 \rightarrow Pos.\ review \end{aligned}, \quad (1)$$

where *Neg.a.g.* – negative mention about some aspect group in review, *Pos.a.g.*– positive mention about some aspect group in review, *Pos. review* – positive review, *Neg. review* – negative review.

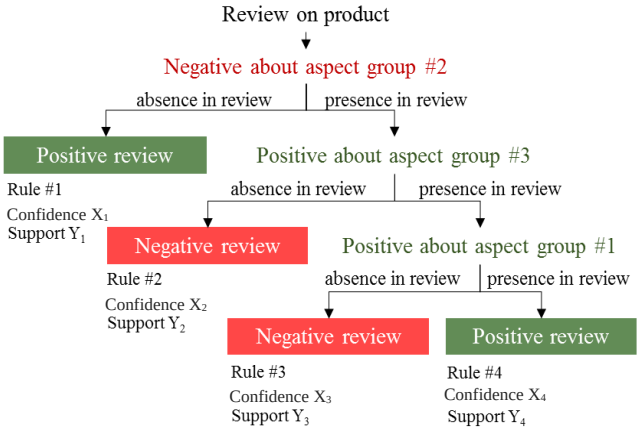


Figure 8. Example of the Decision Tree model

Each rule is characterized by measures of confidence and support. The confidence shows what percentage of reviews containing conditions of some rule, has the same sentiment corresponding to this rule. The support shows percentage of reviews that contain conditions of some rule regarding the entire number of reviews.

E. *Measures for customer satisfaction*

As a measure of the customer satisfaction with product is used a ratio of positive reviews to the sum of positive and negative reviews. The score of customer satisfaction *CS* by product is calculated by (2):

$$CS = \frac{Z^{pos}}{Z^{pos} + Z^{neg}} \cdot 100\%, \quad (2)$$

where Z^{pos} – the number of positive reviews, Z^{neg} – the number of negative reviews.

As a measure of the customer satisfaction with product's aspect groups is used a ratio of positive sentences with mentions of a product's aspect to the sum of positive and negative sentences with mentions of a product's aspect. The score of customer satisfaction cs_j with *j* product's aspect group is calculated by (3):

$$cs_j = \frac{z_j^{pos}}{z_j^{pos} + z_j^{neg}} \cdot 100\%, \quad (3)$$

where z_j^{pos} – number of positive sentences containing mention about the *j* product's aspect group, z_j^{neg} – number of negative comments containing mention about the *j* product's aspect group. Unlike indices in [3] (which often represent the average subjective values obtained with using rating scales) proposed measures show the ratio of positive / negative reviews to total number of reviews. It gives more clearer for understanding of monitoring results.

Significance of aspects group shows how much the sentiment of a review depends on the aspect group in positive and negative sentences, i.e., significance of product's aspects for customers. Let the number of aspect groups is *g/2*, then the number of independent sentimental variables *g*. According to the methodology described in [31] the equation (4) for calculating the significance of variable *m* is:

$$Sign_m = \frac{\sum_{j=1}^{k_m} \left(E_{m,j} - \sum_{i=1}^{q_{m,j}} E_{m,j,i} \cdot \frac{Q_{m,j,i}}{Q_{m,j}} \right)}{\sum_{l=1}^{g} \sum_{j=1}^{k_l} \left(E_{l,j} - \sum_{i=1}^{q_{l,j}} E_{l,j,i} \cdot \frac{Q_{l,j,i}}{Q_{l,j}} \right)} \cdot 100\%, \quad (4)$$

where k_l – number of nodes that were split by attribute *l*, $E_{l,j}$ – entropy of the parent node, split by attribute *l*, $E_{l,j,i}$ – subsite node for *j*, which was split by attribute *l*, $Q_{l,j}$, $Q_{l,j,i}$ – number of examples in the corresponding nodes, $q_{l,j}$ – number of child nodes for *j* parent node.

V. EXPERIMENT

A. *Qualitative and quantitative surveys for hotels*

The proposed approach can be applied for one language (English or French or German etc.). The approach is sensitive for reviewer's eloquence, command of the language, richness of expression, because we don't use specific techniques for this. Nevertheless, we assume that the sensitivity to linguistic peculiarities of review's text will decline with an increase in the training sample.

Efficacy evaluation of the developed IDSS was performed on the data obtained from 635,824 reviews about hotels in the Russian language. The reviews have been collected from the popular Internet resource tophotels.ru for the period of 2003-2013. The initial structure of the collected data consisted of the following fields: hotel name; country name; resort name; visit date; review's text; author's ratings of placement, food, and service. The data was preprocessed and loaded into the database SQL Server 2012.

Classifying of overall sentiment about product used a binary scale (negative and positive). A training set of positive and negative reviews was formed using the collected data on an author's ratings of placement, food, and service. The review site tophotels.ru uses a five-point grading scale. A review can have a maximum total rating

of 15 points, and minimum total rating of 3 points. The training set included 15,790 negative reviews that have 3 and 4 total points, and 15,790 positive reviews that have 15 total points. We did not use the author's ratings for further data processing. Classification of another 604,244 reviews was carried out using a trained classifier.

For the purpose of training an effective sentiment classifier, the accuracy of classification was evaluated for machine learning methods and some peculiarities of their realization (see Table I). The measure accuracy as a ratio of the number of correctly classified reviews to total number of reviews was used to estimate classification accuracy. Accuracy estimation was performed on two sets of data. The first set (Test No. 1) represented a training set of strong positive (15,790) and strong negative reviews (15,790). Classifiers were tested by using cross validation by dividing the first set into 10 parts. The second set (Test No. 2) included random reviews from initial set of reviews (635,824) with different total points (3-15 points) and was labeled manual (497 positive and 126 negative). It was used only for accuracy control of the classifier that had been trained on the first data set.

TABLE I. COMPARISON OF METHODS FOR SENTIMENT CLASSIFICATION

#	Machine learning methods	Vector	Accuracy		
			Test No. 1	*Test No. 2*	*Base line*
1	SVM (linear kernel)[b]	Frequency	94.2%	83.1%	72.8%
2	SVM (linear kernel)	Binary	95.7%	84.1%	82.9%
3	NB[a]	Binary	96.1%	83.7%	81%
4	NB	Frequency	97.6%	92.6%	78.7%
5	NB (stop-words)	Frequency	97.7%	92.7%	-
6	Bagging NB	Frequency	97.6%	92.8%	-
7	NB ("negations")	Frequency	98.1%	93.6%	-

a. Naïve Bayes Classifier. b. Support Vector Machine

To estimate influence of negative particles "not" and "no", the tagging technique was used; for example, the phrase "not good" was marked as "not_good", and was regarded by the classifier as one word. This negation technique allowed the increasing of sentiment classification accuracy (see classifier #7). Accuracy values are presented in the Table I. We used results obtained by Pang and Lee [19] for movie-review domain as a base line. Our results of classification accuracy looks better than the base line. This can be explained that we used more large training sample (31,580 vs 2,053). The most efficient ML method was Naïve Bayes classifier with negation technique (classifier #7). In Figure 9 are presented ROC-curves classifiers #4 and #7. The classifier #7 was trained on the training set and was used for further sentiment analysis of unlabeled reviews.

Using the aspect extraction algorithm (Section III), we extracted the nouns that were divided into seven basic aspect groups (see Figure 10): "beach/swimming pool", "food", "entertainment", "place", "room", "service", "transport". The following step was extracting and sentiment classification sentences with words from aspect groups using classifier #7. However, not all sentences with aspects had a clearly expressed sentiment; therefore, the sentences with poorly expressed sentiment were filtered out using threshold *h*. Threshold *h* was chosen empirically and it allowed to cut sentences with weak sentiment. This algorithm of aspect-based sentiment analysis (Figure 6) is very simple and primitive, but it allows to realize its main aim - to identify sentences with strong sentiments (that form overall sentiment of review) with mentions about product's aspects, and filter out sentences with weak sentiment or without mentions about product's aspects.

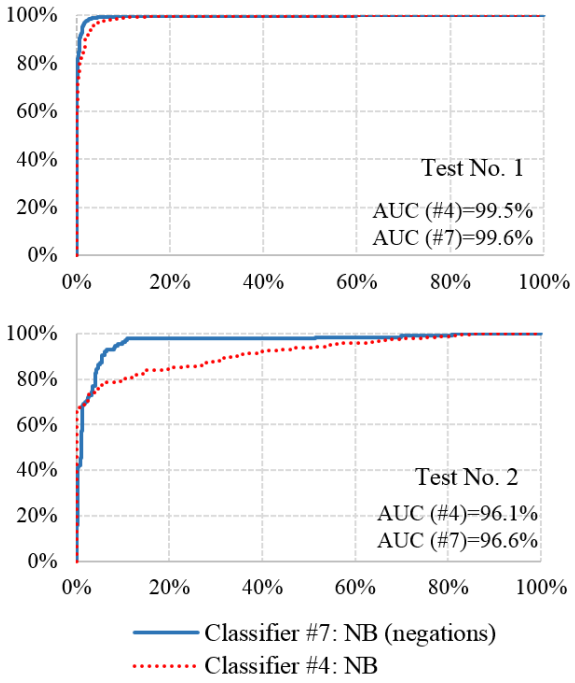


Figure 9. Comparison of ROC-curves for classifiers #7 and #4

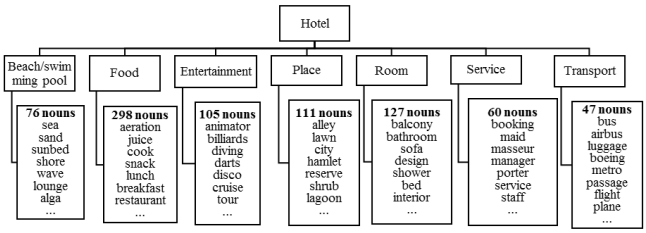


Figure 10. Aspect groups of object "hotel"

In present work we give an example of qualitative and quantitative surveys for two 5-star hotels "A" (1692 reviews) and "B" (1,300 reviews) located on the resort Sharm el-Sheikh (63,472 reviews) in Egypt. Firstly, we will make a quantitative survey, measure customer satisfaction, compare it with average satisfaction in the whole resort, detect negative trends by each hotel's aspect group, and identify problems in the quality of hotels.

The dynamics of customer satisfaction calculated by equation (2) is presented in Figure 11. Concerning the hotel "A", there is a positive upward satisfaction trend from 2009, and it fixes on the average-resort level in 2013. Concerning the hotel "B", in 2012 there was a sharp satisfaction decline and the same sharp increase in 2013. We can also notice that on a monthly graph (Figure 12). For the hotel "B", satisfaction decrease started in June 2012, and stopped in October 2012. Then, customer satisfaction grew to the level that was higher than the average resort level being ahead of its competitor – hotel "A".

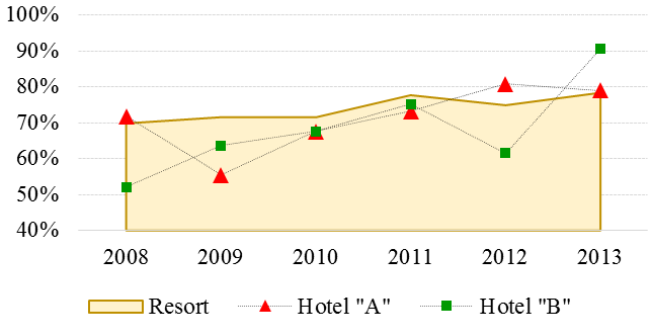


Figure 11. Dynamics of the consumer satisfaction by years.

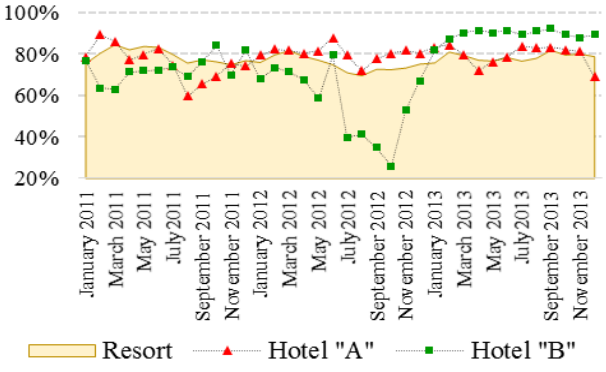


Figure 12. Dynamics of the customer satisfaction by months.

To find reasons of hotel "B" satisfaction decrease, we will examine structure of customer satisfaction with product in Figure 13. We can see that in 2012, the hotel "B" on average was second to the hotel "A" in such aspects as "room" (Δ12%), "place" (Δ8%), "service" (Δ5%), "beach/swimming pool" (Δ3%) and "entertainment" (Δ3%). To complement the overall picture of the level of quality we have included in the analysis mentions about "theft" and "intoxication". In 2012, the Hotel "B" had more registered cases of intoxication in September 2012, as well as cases of theft in August 2012 (see Figure 15). We should also note that one of the reasons of customer dissatisfaction with the hotel "B" was the initiated repair of hotel rooms and buildings, which, however, paid off in 2013. Customer satisfaction with the hotel "A" aspects conforms with the average resort level.

In 2013, customer satisfaction with product's aspect groups with the hotel "B" exceeded the average level in all aspect groups (see Figure 14). Customer satisfaction with the hotel "A" dropped lower than average values in such aspects as "service" (Δ3%), "food" (Δ3%), "beach/swimming pool" (Δ3%) and "transport" (Δ4%). For hotel "A" manager arise questions like: which aspects are the most significant for the customer and that should be improved in the first place, is it possible to "substitute" the dissatisfaction with the service, e.g., by tasty food or employ new entertainer?

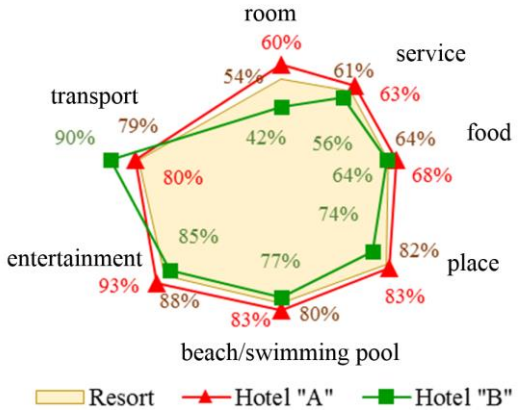


Figure 13. Comparison of the structure of customer satisfaction in 2012.

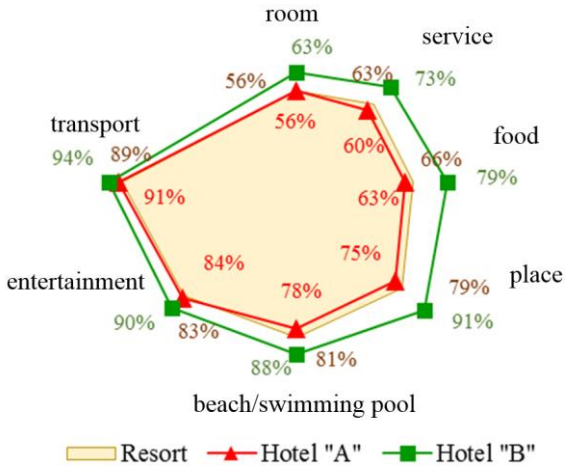


Figure 14. Comparison of the structure customer satisfaction in 2013.

Decision Trees were constructed using algorithm C4.5 and a tool "Deductor" [31]. At the first step was constructed a tree for the all hotels of resort. Extracted rules are represented in Table III. At the second step trees were constructed for hotel "A" and hotel "B". Significance values of a product's aspect groups are represented in Table II. In Figure 13, there are the decision trees created for hotel "A" and hotel "B". Due to the large size of the produced decision tree of the whole resort we omitted it, but in the Table III its rules are presented that have confidence >80% and support >5% (it is five rules from 27).

By analyzing significance values (see Table II) of product's aspect group on overall review's sentiment (i.e., on customer satisfaction), we can say that the main factors of consumer dissatisfaction are a low service level (34.8%), problems with food (16%), and complaints about the hotel rooms (4%). The most critical aspect group for the hotel "B" is "room" (57.3%). In absence of negative opinions on the aspect group "room", the review would be positive with a confidence of 95.5% (see Table III, rule #10). That is why the repair that was performed facilitated to a significant increase of consumer satisfaction. The most critical aspect group for the Hotel "A" is "service" that corresponds with the resort in a whole.

	Hotel "A"									Hotel "B"								
	beach/swimming pool	food	entertainment	place	room	service	transport	intoxication	theft	beach/swimming pool	food	entertainment	place	room	service	transport	intoxication	theft
January-11	91%	65%	87%	84%	55%	60%	90%	1%	6%	86%	75%	85%	78%	57%	67%	73%	3%	0%
February-11	95%	82%	94%	75%	61%	80%	45%	0%	3%	93%	87%	43%	39%	28%	33%	37%	1%	0%
March-11	76%	70%	88%	88%	60%	78%	73%	0%	2%	81%	77%	55%	70%	31%	42%	52%	1%	0%
April-11	82%	67%	87%	81%	62%	65%	74%	0%	1%	82%	78%	70%	76%	42%	57%	68%	0%	0%
May-11	89%	68%	87%	81%	58%	57%	87%	2%	2%	82%	73%	77%	77%	44%	64%	71%	0%	0%
June-11	85%	77%	93%	78%	50%	62%	77%	1%	8%	84%	72%	83%	75%	48%	60%	79%	2%	0%
July-11	74%	68%	80%	64%	38%	38%	88%	5%	4%	81%	72%	87%	74%	46%	59%	84%	1%	0%
August-11	82%	62%	79%	67%	40%	41%	44%	7%	10%	76%	66%	87%	71%	43%	57%	85%	0%	3%
September-11	77%	58%	89%	70%	45%	47%	72%	8%	5%	86%	68%	91%	86%	50%	61%	86%	2%	1%
October-11	78%	59%	93%	80%	48%	54%	76%	5%	5%	89%	71%	91%	85%	56%	75%	93%	1%	1%
November-11	81%	63%	93%	80%	55%	65%	88%	4%	2%	79%	66%	84%	78%	47%	57%	90%	2%	2%
December-11	79%	63%	88%	87%	49%	62%	94%	4%	3%	88%	73%	87%	81%	49%	63%	89%	1%	3%
January-12	86%	64%	89%	86%	54%	61%	88%	3%	3%	84%	66%	86%	78%	42%	56%	95%	3%	3%
February-12	86%	64%	95%	87%	59%	63%	88%	2%	1%	84%	70%	87%	80%	46%	63%	94%	1%	2%
March-12	85%	71%	97%	90%	64%	68%	94%	3%	2%	88%	65%	89%	76%	39%	66%	97%	8%	5%
April-12	84%	69%	91%	86%	58%	64%	76%	3%	2%	81%	62%	85%	79%	40%	59%	93%	4%	2%
May-12	85%	70%	94%	86%	61%	67%	77%	5%	2%	77%	58%	71%	71%	32%	52%	80%	7%	1%
June-12	86%	69%	95%	88%	62%	69%	76%	4%	4%	88%	79%	86%	77%	52%	59%	90%	4%	1%
July-12	84%	67%	92%	85%	59%	60%	88%	5%	4%	69%	75%	93%	51%	26%	55%	45%	2%	0%
August-12	81%	64%	92%	81%	55%	57%	86%	5%	3%	62%	63%	80%	63%	30%	45%	56%	1%	7%
September-12	84%	70%	91%	79%	58%	68%	71%	5%	1%	63%	60%	73%	63%	33%	45%	78%	15%	4%
October-12	83%	68%	90%	81%	57%	63%	82%	3%	1%	62%	48%	87%	56%	35%	35%	72%	7%	2%
November-12	79%	66%	92%	83%	59%	62%	74%	2%	0%	75%	60%	93%	70%	47%	67%	86%	4%	1%
December-12	77%	66%	93%	80%	60%	62%	77%	2%	0%	83%	70%	92%	73%	53%	65%	85%	5%	0%
January-13	81%	68%	94%	82%	61%	64%	75%	3%	1%	92%	75%	94%	78%	57%	70%	92%	2%	4%
February-13	82%	66%	95%	85%	62%	64%	87%	1%	3%	88%	82%	95%	88%	61%	75%	96%	1%	2%
March-13	78%	61%	94%	81%	57%	57%	77%	1%	1%	93%	86%	95%	89%	66%	79%	98%	2%	1%
April-13	75%	58%	91%	82%	52%	57%	89%	3%	2%	93%	89%	95%	90%	74%	82%	99%	1%	0%
May-13	76%	59%	91%	83%	52%	59%	80%	2%	4%	86%	81%	92%	88%	67%	74%	97%	0%	0%
June-13	77%	62%	90%	81%	53%	59%	65%	2%	2%	88%	80%	95%	88%	62%	71%	91%	3%	1%
July-13	77%	66%	91%	83%	58%	63%	72%	5%	3%	89%	76%	94%	87%	61%	69%	89%	3%	3%
August-13	75%	67%	92%	89%	59%	63%	76%	2%	1%	87%	79%	92%	91%	63%	76%	94%	2%	2%
September-13	79%	67%	88%	94%	54%	57%	88%	1%	1%	84%	78%	94%	86%	65%	76%	86%	2%	2%
October-13	74%	64%	83%	88%	53%	50%	44%	7%	0%	88%	77%	95%	88%	59%	71%	85%	1%	3%
November-13	77%	63%	86%	86%	56%	57%	62%	3%	0%	89%	76%	94%	93%	59%	70%	90%	2%	2%
December-13	80%	63%	83%	88%	52%	58%	41%	4%	0%	86%	78%	94%	95%	61%	71%	91%	1%	1%

Figure 15. Customer satisfaction with product's aspect groups by month.

Using significance values, we can relate each aspect group with Kano's model categories [19]. Negative mentions about aspect group "service" in review have high significance on customer satisfaction, but positive mentions have significance is near zero (34.8% vs. 0.7%). That's why we can say that aspect group "service" relates to "Must-be quality" of Kano's categories. This is interpreted as a positive mentions about aspect groups "service" do not have an influence on sentiment of review, i.e., on overall satisfaction with hotel. That means the consumer a priori awaits a high-level service as a matter of course.

TABLE II. SIGNIFICANCE OF PRODUCT'S ASPECT GROUPS ON OVERALL

Aspect group	Kano's model category	Sentiment of mention	Significance values		
			Resort	*Hotel "A"*	*Hotel "B"*
Service	Must-be quality	Negative	34.8%	60.2%	-
		Positive	0.7%	-	-
Food	One-dimensional quality	Negative	30.3%	27.2%	30.3%
		Positive	16%	-	-
Entertainment	Attractive quality	Negative	-	-	-
		Positive	8.5%	12.7%	12.4%
Room	One-dimensional quality	Negative	4%	-	57.3%
		Positive	2.1%	-	-
Beach/swimming pool	Attractive quality	Negative	0.2%	-	-
		Positive	2.5%	-	-
Place	Attractive quality	Negative	-	-	-
		Positive	1%	-	-
Transport	Indifferent quality	Negative	-	-	-
		Positive	-	-	-

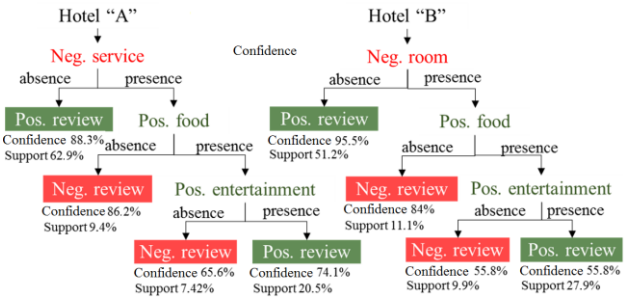


Figure 16. Decision trees for hotels.

For such aspect groups as "food" and "room" significance values are comparable (30.3% vs. 16% for "food" and 4% vs. 2.1% for "room"), that relates to "One-dimensional quality" of Kano's categories. That means that customer satisfaction rising with rising quality of aspect groups "food" and "room" and reducing when reducing quality of these aspect groups.

Aspect groups "beach/swimming pool", "entertainment" and "place" relates to "Attractive quality" because only positive mentions about them have a significance on customer satisfaction. That means that customer satisfaction rising with rising quality of aspect groups "beach/swimming pool", "place" and "territory", but does not reducing when reducing quality of these aspect groups.

In some cases, positive mentions about "food" and "entertainment" simultaneously in a review could substitute negative mentions about "services" and provide a positive review. That is why the hotel's aspects, which are contributing to customer satisfaction and important for both the resort and for the hotels, are good food and amusing entertainment activities. Customer satisfaction with these aspect groups can overlap dissatisfaction with

"service" or "rooms" and make the customer overall satisfied (see Table III, rules #5, #7, #11).

TABLE III. RULES EXTRACTED BY USING DECISION TREES

#	Rules	S[a]	C[b]
	Extracted rules on resort reviews		
1	$Food^{+} \cap \overline{Service^{-}} \cap \overline{Food^{-}}$ = Positive review	37.2%	97.4%
2	$Food^{+} \cap \overline{Service^{-}} \cap Food^{-} \cap Beach^{+}$ = Positive review	11%	86.2%
3	$\overline{Food^{+}} \cap \overline{Service^{-}} \cap \overline{Food^{-}} \cap \overline{Room^{-}}$ = Positive review	10.6%	83.9%
4	$\overline{Food^{+}} \cap Service^{-} \cap \overline{Entertainment^{+}}$ = Negative review	6.9%	92.3%
5	$Food^{+} \cap Service^{-} \cap \overline{Food^{-}} \cap Entertainment^{+}$ = Positive review	5.8%	88.4%
	Extracted rules on Hotel "A" reviews		
6	$\overline{Service^{-}}$ = Positive review	62.9%	88.3%
7	$Food^{+} \cap Service^{-} \cap Entertainment^{+}$ = Positive review	20.5%	74.1%
8	$\overline{Food^{+}} \cap \overline{Service^{-}}$ = Negative review	9.4%	86.2%
9	$Food^{+} \cap Service^{-} \cap \overline{Entertainment^{+}}$ = Negative review	7.2%	65.6%
	Extracted rules on Hotel "B" reviews		
10	$\overline{Room^{-}}$ = Positive review	51.2%	95.5%
11	$Food^{+} \cap Room^{-} \cap Entertainment^{+}$ = Positive review	27.9%	81%
12	$\overline{Food^{+}} \cap Room^{-}$ = Negative review	11.1%	84%
13	$Food^{+} \cap Room^{-} \cap \overline{Entertainment^{+}}$ = Negative review	9.9%	55.8%

a. Support. b. Confidence

Then we compared frequency of nouns from aspect groups and significance values of aspect groups (see Table IV). As we can see, significance values do not correlate with frequency of nouns. For example, the most frequent aspect groups in reviews are "room" and "beach/swimming pool", but these aspect groups have low significance in contrast to "service" and "food". It also confirms the correctness of the chosen approach to the estimation of significance values of aspect groups.

TABLE IV. COMPARISON OF FREQUENCY OF NOUNS AND THEIR SIGNIFICANCE VALUES

Aspect groups	Number of nouns in aspect group	Frequency of nouns from aspect group	Significance values of aspect groups	
			Negative	Positive
Service	60	13.7%	34.8%	0.7%
Food	298	15.0%	30.3%	16.0%
Entertainment	105	8.9%	-	8.5%
Room	127	23.3%	4.0%	2.1%
Beach/swimming pool	76	25.1%	0.2%	2.5%
Place	111	8.8%	-	1.0%
Transport	47	5.1%	-	-

The performed qualitative survey allowed the detection of the main ways to increase customer satisfaction for hotel "A". The problem aspect groups identified through quantitative survey correspond to the most significant aspects detected during the qualitative research. Hotel "A" manager should firstly increase "service" quality, and then increase the quality of "food" and "beach/swimming pool" maintenance. "Transport" problems – concerning flights, early check-in, and baggage storage – are not significant for customers and can be solved in the frames of service improvement. The process of service quality increase can take much time; that is why organizing entertainment and animated programs together with enhancement of restaurant service could be immediate measures for increasing customer satisfaction. Specification of managerial decisions can be performed on the basis of the information on existing problems contained in negative reviews. The extracted sentences on aspects can be directed to the appropriate hotel services.

B. Qualitative and quantitative surveys for banks

For the qualitative and quantitative surveys for banks, we used small sample of consumer reviews – 1,153 reviews of Russian banks from site banki.ru. The sample consists of 304 positive reviews and 849 negative reviews in Russian.

It should be noted that the data is skewed towards negative reviews. This is because customers often leave a review in the case of dissatisfaction with the bank. In the absence of dissatisfaction, customers do not have a motive to leave a positive review. As we see studies in the respect of hotels and resorts, in contrast, show a skew towards positive reviews. We should note that identifying all of a product's aspects depends on reviewers' mentioning them all in the reviews. That is why very important to make satisfaction customer research using a large enough sample of positive and negative reviews, which would have covered mentions with all aspects of the product. In this case, the observing skew of positive and negative reviews has no effect on the significance values (4) of aspects and identified rules of decision trees (1) as for any other data mining problem. On the other hand, the skew is important for measuring customer satisfaction with product (2) and product's aspects (3).

To estimate accuracy of machine learning methods we used a cross-validation. Evaluation of the accuracy of the classifiers is calculated as the proportion of correctly classified positive and negative feedback on their total number. Results are presented in the Table V.

TABLE V. COMPARISON OF METHODS FOR SENTIMENT CLASSIFICATION

#	Machine learning methods	Vector	Accuracy
1	NB[a]	Binary	86.5%
2	NB	Frequency	86.8%
3	SVM (linear kernel)[b]	Binary	87.7%
4	SVM (linear kernel)	Frequency	85.0%
5	NB ("negations")	Frequency	88.0%

a. Naïve Bayes Classifier. b. Support Vector Machine

Accuracy values of the classification machine learning methods are comparable. For the sentiment analysis was chose Naive Bayes classifier with frequency vector. In this experiment, we also used tagging technique of negative particles "not" and "no". This technique has improved the classification accuracy to 88%. On the sample of bank reviews were extracted various aspects of banking.

Extracted aspects grouped by groups are shown in the Table VI. These aspect groups are used for the aspect-based sentiment analysis.

TABLE VI. ASPECT GROUPS OF OBJECT "BANK"

Aspect groups	Aspect nouns
Staff	Administrator, manager, supervisor, expertise, consultant, incompetence, maintenance, operator, guard, staff, employee, management, employees, specialist.
Credit	Profile, loan, debt, borrower, application, statement, mortgage, credit, lending limits, approval, waiver, redemption, delay, consideration.
Deposit	Contributor, deposit, contribution.
Card	Lock, ATM, release, holder, card, credit card, reissue.
Settlement and cash services (SCS)	Cashier, receipt, commission, transfer, assignment, debit, bill, terminal transaction.

Then we carried out quantitative and qualitative surveys of consumer satisfaction for four Russian banks: VTB24 (120 reviews), Alpha-Bank (131 reviews), Sberbank (232 reviews), Bank Russian Standard (86 reviews). The results of the quantitative survey are presented in the Table VII. Among these banks, the most satisfied customers are the customers of the VTB24 (35% positive reviews). Then follows Alfa-Bank and Sberbank (26.7% and 19.8%, respectively). At the last place is the Bank Russian Standard (14% positive reviews). These satisfaction values corresponds with customers' satisfaction values presented on the site as known as "People's rating".

TABLE VII. THE CUSTOMER SATISFACTION WITH ASPECT GROUPS AND OVERALL

	VTB24	Alpha-Bank	Sberbank	Bank Russian Standard
Staff	54%	48%	42%	38%
Credit	40%	25%	29%	28%
Deposit	100%	100%	69%	50%
Card	41%	48%	39%	29%
Settlement and cash services	42%	40%	31%	29%
Overall satisfaction	**35.0%**	**26.7%**	**19.8%**	**14.0%**
People's rating[a]	**38.6**	**37.1**	**32.7**	**31.6**

a. Ratings of banks, based on the assessments of customers, exposed on the Internet site banki.ru

Aspect-based sentiment analysis allowed evaluating consumer satisfaction with specific aspect groups of banking services. For example, although the first place in overall satisfaction, VTB 24 has a lower satisfaction on card products than those of Alfa Bank. However, there are problems with credit products at Alfa-Bank, because customers complain about unreasonable delay cases and the emergence of debt on loans.

Based on sentiment analysis reviews and aspect-based sentiment analysis we constructed decision tree. Obtained decision rules with confidence of more than 75% are shown in Table VIII. Rule #1 means that the absence of positive mentions about the staff and cards lead to negative reviews with confidence 92.9%. 60.2% of all reviews contain this rule. Positive mentions about staff without negative mentions about staff and settlement and cash services lead to positive reviews with confidence 79.4% (rule #2).

TABLE VIII. RULES EXTRACTED BY USING DECISION TREES

#	Rules	S[a]	C[b]
1	$\overline{Staff^+} \cap \overline{Card^+} = Negative\ review$	60.2%	92.9%
2	$Staff^+ \cap \overline{Staff^-} \cap \overline{SCS^-} = Positive\ review$	20.2%	79.4%
3	$Staff^+ \cap Staff^- = Negative\ review$	7.4%	85.9%
4	$\overline{Staff^+} \cap Card^+ \cap Card^- = Negative\ review$	4.1%	87.2%

a. Support. b. Confidence

Table IX shows the estimated values of the significance values of aspect groups for customers. The greatest impact on customer satisfaction have positive and negative mentions about bank' staff. Next in significance aspect groups are card products and Settlement and cash services. The rest of the aspect groups have significance values near zero.

TABLE IX. SIGNIFICANCE OF PRODUCT'S ASPECT GROUPS

Aspect groups	Kano's model category	Sentiment of mention	
		Negative	Positive
Staff	One-dimensional quality	19.5%	54.7%
Credit	Indifferent quality	0%	0%
Deposit	Indifferent quality	0%	0%
Card	One-dimensional quality	6.3%	14.9%
Settlement and cash services	Must-be quality	4.6%	0%

VI. CONCLUSION AND FUTURE WORK

Poor quality of products and services contributes to a decrease of customer satisfaction. On the other hand, under the conditions of stiff competition, there are no barriers for the consumer to change the supplier of goods and services. All these things can cause loss of clients and a decrease of a company's efficiency indexes. Therefore, maintaining high-quality standards should be provided by effective managerial decisions and based on opinion mining as a feedback.

The suggested conception of decision support based on the developed approach of text data processing and analysis allows performing quantitative and qualitative surveys of customer satisfaction using computer-aided procedures, and making effective managerial decisions on product quality management. The present conception allows effective reduction of labor intensity of customer satisfaction research that makes it available for use by a wide range of companies.

A prototype of IDSS was developed on the basis of the suggested conception. The performed experiment has proved its efficacy for solving real problems of quality management and consistency of the results obtained. IDSS enables companies to make decisions on quality control based on analytical processing of text data containing implicit information on client satisfaction.

Future research on the given topic can be devoted to automatic annotating of text data, representing text amount of review in the form of a summary, and extracting useful and unique information.

REFERENCES

[1] N. Yussupova, M. Boyko, D. Bogdanova, and A. Hilbert, "A Decision Support Approach for Quality Management based on Artificial Intelligence Applications," Proceedings of The Third International Conference on Data Analytics (DATA ANALYTICS 2014), Rome, Italy, 2014, pp. 112-121, ISBN: 978-1-61208-358-2.

[2] ISO 9000-2008 The quality management system. Fundamentals and vocabulary.

[3] ISO10004:2010 Quality management. Customer satisfaction. Guidelines for monitoring and measuring.

[4] D. Gräbner, M. Zanker, G. Fliedl, and M. Fuchs, "Classification of Customer Reviews based on Sentiment Analysis," Proceedings of the International Conference in Helsingborg, Springer Vienna, Jan. 2012, pp. 460-470, ISBN 978-3-7091-1141-3.

[5] M. Taboada, J. Brooke, M. Tofiloski, K. D. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," Computational Linguistics, vol. 37(2), pp. 267-307, June 2011, doi:10.1162/COLI_a_00049.

[6] Y. Jo and A. Oh, "Aspect and Sentiment Unification Model for Online Review Analysis," Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11), ACM New York, Feb. 2011, pp. 815-824, ISBN: 978-1-4503-0493-1.

[7] B. Lu, M. Ott, C. Cardie, and B. Tsou, "Multi-aspect Analysis with Topic Models," Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW '11), Dec. 2011, pp. 81-88, ISBN: 978-0-7695-4409-0.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol. 3 (4–5), pp. 993-1022, Jan. 2003, doi:10.1162/jmlr.2003.3.4-5.993.

[9] B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," Proceedings of the 14th international conference on World Wide Web (WWW '05), ACM New York, May 2005, pp. 342-351, ISBN: 1-59593-046-9.

[10] W. Kasper and M. Vela, "Sentiment Analysis for Hotel Reviews," Proceedings of the Computational Linguistics-Applications Conference (CLA-2011), Oct. 2011, pp. 45–52.

[11] G. Ganu, A. Marian, and N. Elhadad, "URSA – User Review Structure Analysis: Understanding Online Reviewing Trends," Rutgers DCS Technical Report No. 668, April 2010. [Online]. Available from: http://spidr-ursa.rutgers.edu/resources/TR_LRE.pdf.

[12] S. Blair-Goldensohn, K. Hannan, and R. McDonald, "Bilding a Sentiment Summarizer for Local Service Reviews," Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09), March 2009, pp. 514-522.

[13] E. Bjørkelund, T.Burnett, K. Nørvåg, "A Study of Opinion Mining and Visualization of Hotel Reviews," Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services (IIWAS 2012), ACM New York, Dec. 2012, pp. 229-238, ISBN: 978-1-4503-1306-3.

[14] J. Ajmera, H. Ahn, M. Nagarajan, A. Verma, D. Contractor, S. Dill, and M. Denesuk, "A CRM system for Social Media," Proceedings of the 22nd international conference on World Wide Web (WWW '13), May 2013, pp. 49-58, ISBN: 978-1-4503-2035-1.

[15] M. Bank, "AIM – A Social Media Monitoring System for Quality Engineering," PhD thesis, Universität Leipzig, June 2013, p. 235.

[16] H. Wang, Y. Lu, and C. Zhai, "Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach," Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10), ACM New York, July 2010, pp. 783-792, ISBN: 978-1-4503-0055-1.

[17] J. C. de Albornoz, L. Plaza, P. Gervás, and A. Díaz, "A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating," Proceedings of the 33rd European Conference on Information Retrieval (ECIR '11), April 2011, pp. 55-66, ISBN: 978-3-642-20160-8.

[18] H. Wachsmuth, M. Trenkmann, B. Stein, G. Engles, and T. Palakarska, "A Review Corpus for Argumentation Analysis," Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing, Springer (Kathmandu, Nepal), LNCS, April 2014, pp. 115-127, ISBN 978-3-642-54902-1.

[19] J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua, "Aspect Ranking: Identifying Important Product's aspects from Online Consumer Reviews," Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11), June 2011, pp. 1496-1505, ISBN: 978-1-932432-87-9.

[20] N. Kano, N. Seraku, F. Takashi, and S. Tsuji, "Attractive quality and must-be quality," In The Journal of the Japanese Society for Quality Control, 1984, V. 14, pp. 39-48.

[21] J. Thomsen, E. Ernst, C. Brabrand, and M. Schwartzbach, "WebSelF: A Web Scraping Framework," Proceedings of the 12th international conference on Web Engineering (ICWE 2012), July 2012, pp. 347-361, ISBN: 978-3-642-31752-1.

[22] R. Penman, T. Baldwin, and D. Martinez, "Web scraping made simple with sitescraper," 2009. [Online]. Available from: http://sitescraper.googlecode.com.

[23] J. Yi, T. Nasukawa, W. Niblack, and R. Bunescu, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," In Proceedings of the 3rd IEEE international conference on data mining, ICDM 2003, pp. 427-434.

[24] A. G. Pazelskaya and A. N. Soloviev, "Method of the determination emotions in the lyrics in Russian," Computer program linguistics and intellectual technologies, Issue 10 (17), 2011, pp. 510-522.

[25] B. Pang and L. Lee, "Thumbs up? Sentiment Classification using Machine Learning Techniques," Proceedings of the Conference on Empirical Methods in Natural. Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.

[26] C. Manning, P. Raghavan, and H. Schuetze, "An Introduction to Information Retrieval," Cambridge University Press. Cammbridge, England, 2009, pp. 1-544.

[27] N. Yussupova, D. Bogdanova, and M. Boyko, "Algorithms and software for sentiment analysis of text messages using

machine learning," Vestnik USATU, T. 16-6(51), 2012, pp. 91-99.

[28] N. Yussupova, D. Bogdanova, and M. Boyko, "Applying of Sentiment Analysis for Texts in Russian Based on Machine Learning Approach," Proceedings of the 2nd International Conference on Advances in Information Mining and Management (IMMM2012), Venice, Italy, 2012, pp. 8-14. ISBN: 978-1-61208-227-1.

[29] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, "Red Opal: Product-Feature Scoring from Reviews," Proceedings of the 8th ACM conference on Electronic commerce, 2007, pp. 182-191. ISBN: 978-1-59593-653-0.

[30] O. N. Ljashevskaja and S. A. Sharov, "The new frequency vocabulary of Russian lexic based on the Russian National Corpus," RAS, Institut of Russian language named of V. V. Vinogradov, 2009. [Online]. Available from: http://dict.ruslang.ru/freq.php.

[31] Deductor. The Algorithm Manual (ver. 5.2.0), BaseGroup Labs, 2010. [Online]. Available from: http://www.basegroup.ru/download/guide_algorithm_5.2.0.pdf

Human Activity Recognition using a Semantic Ontology-Based Framework

Rosario Culmone
Computer Science Division,
School of Sciences and Technologies,
University of Camerino
Email: rosario.culmone@unicam.it

Paolo Giuliodori
Computer Science Division,
School of Sciences and Technologies,
University of Camerino
Email: paolo.giuliodori@unicam.it

Michela Quadrini
Computer Science Division,
School of Sciences and Technologies,
University of Camerino
Email: michela.quadrini@unicam.it

***Abstract*—In the last years, the extensive use of smart objects embedded in the physical world, in order to monitor and record physical or environmental conditions, has increased rapidly. In this scenario, heterogeneous devices are connected together into a network. Data generated from such system are usually stored in a database, which often shows a lack of semantic information and relationship among devices. Moreover, this set can be incomplete, unreliable, incorrect and noisy. So, it turns out to be important both the integration of information and the interoperability of applications. For this reason, ontologies are becoming widely used to describe the domain and achieve efficient interoperability of information system. An example of the described situation could be represented by Ambient Assisted Living context, which intends to enable older or disabled people to remain living independently longer in their own house. In this contest, human activity recognition plays a main role because it could be considered as starting point to facilitate assistance and care for elderly. Due to the nature of human behavior, it is necessary to manage the time and spatial restrictions. So, we propose a framework that implements a novel methodology based on the integration of an ontology for representing contextual knowledge and a Complex Event Processing engine for supporting timed reasoning. Moreover, it is an infrastructure where knowledge, organized in conceptual spaces (based on its meaning) can be semantically queried, discovered, and shared across applications. In our framework, benefits deriving from the implementation of a domain ontology are exploited into different levels of abstraction. Thereafter, reasoning techniques represent a preprocessing method to prepare data for the final temporal analysis. The results, presented in this paper, have been obtained applying the methodology into AALISABETH, an Ambient Assisted Living project aimed to monitor the lifestyle of old people, not suffering from major chronic diseases or severe disabilities.**

Keywords–Pattern Recognition; OntoAALISABETH Domain Ontology; Semantic Reasoning; Complex Event Processing (CEP).

I. INTRODUCTION

In the last years, the extensive use of smart objects embedded in the physical world in order to obtain information has increased rapidly. In other words, such sensor network allows to monitor and record physical or environmental conditions, especially interactions of users with the physical world. In order to reach this aim, the network is composed of a large and heterogeneous sources. The issue is that typical technologies for recording data do not allow to describe the relation of a sensor with the network. Furthermore, data can be incomplete, unreliable, incorrect, and could happen that one type of information is expressed by using different type of physical measures. So, to process these data generated by several heterogeneous sources, it turns out to be important both the integration of information and the interoperability of applications. In such scenario, these data are stored in a data repository, usually a Database (DB), which often shows a lack of semantic information and relationships among components of the system. So, acquired data from smart objects need to be treated according to their semantics. For this reason, in order to successfully monitor a situation, it is typically necessary to integrate stored data with static data and background knowledge. Ontologies represent a tool to connect these aspect. In fact, they provide a shared understanding of a domain, hence allowing semantic interoperability. This is the approach that we have presented into the initial proceeding [1].

An example of the described situation could be represented by an Ambient Assisted Living (AAL) context. The AAL program, promoted by the European Commission [2], intends to enable older or disabled people for the purpose of remaining living independently longer in their own house with an improved quality of life [3][4]. So, in the user domestic environment a wide network of smart objects is installed, whose task is to provide the possibility to monitor the user lifestyle. In order to reach this aim, the smart home (SH) relies on many different types of objects: from clinical devices for the user's health to indicators of presence, from temperature and humidity measurements to fridge and door opening sensors.

In this context, also at national and regional level [5], there are many on-going experiments among which AALISABETH (Ambient-Aware LIfeStyle tutoring for A BETter Health) [6], a project running in Regione Marche. This project has the objective to develop a new technology, based on the use of non-invasive sensor network, for monitoring the lifestyle of old people (65+), not suffering from major chronic diseases or severe disabilities. In particular, the main goal of this project is to detect a set of abnormal behaviors that could bring to the onset of the most common diseases. In particular, the same activity can acquire different meaning depending on the time of day. In order to reach this goal, a set of sensors has been selected and interconnected through an heterogeneous communicating network that wires the AALISABETH SH. Data collected from such variety of devices, weather environment, wearable or clinical sensors, are store in a proper database. In order to answer to the requirements of previous portrayed project, we develop a novel methodology, which is able to detect particular behaviours, compare them evolving with time, or determine

the order in which events occurred. In situation monitoring, geo-spatial information is also of great importance, since it enables to locate events in the real world.

For these reasons, our methodology integrates an ontology for representing contextual knowledge with rule-based and a Complex Event Processing (CEP) engine for supporting the timed reasoning. It is an infrastructure where the knowledge, organized in conceptual spaces (based on its meaning) can be semantically queried, discovered, and shared across applications. The ontology is introduced because it is able to provide a shared understanding of a domain, hence allowing semantic interoperability. In addition, it has the ability to reuse knowledge and integrate several knowledge domains. Moreover, it is built following a pyramidal structure in order to distinguish two types of knowledge, static and dynamic. The former describes the domain, while the latter models the context acquisition, in particular sensor data, in order to describe the AAL domain, to organize data according to their semantic meaning and to select them during the pre-processing phase. On the other hand, CEP engine has been introduced in the proposed framework due to the expressiveness limitation of ontology, which lacks of temporal reasoning. In fact, traditional methods have not focused on reasoning over time and space, which is necessary to capture some of the important characteristics of streaming data and events. Moreover, the benefits of the framework can be noted in the Section IV, where a concrete case of use is presented. More specifically, the framework has been tested in a standard flat populated by an elderly person that has a regular behaviour.

The paper is structured as follows: Section II examines the related literature concerning the topics addressed in this work. Section III explains the motivation of the proposed methodology, in particular it provides a detailed description of the framework architecture. Section IV is entirely dedicated to implementation of framework with its different components, starting from data source, going trough semantic and ending with pattern recognition. At the end of this section a concrete example is provided, which allows to validate the proposed methodology. Section V contains the conclusion and some possible future development.

II. Related Work

Human activity discovery and recognition play an important role in a wide range of applications in the AAL domain in order to facilitate assistance and care for elderly. Moreover, they represent an active and ambitious research area because of the large amount of noise in data and the difficulty of modelling situations [7]. In addition, each human has different ways to perform the activity, but also people can do several activities at the same time, or different places may be needed to perform a particular activity.

In this scenario, methods to recognise human activity have been widely studied for long time and several approaches have been developed. They can be divided into three main categories: statistical, probabilistic and logic. The former is based on machine learning techniques, including both supervised and unsupervised human behaviour recognition. Moreover, there is a wide range of algorithms and models for human recognition based on statistical approach. For instance, Fleury et al. [7] classify human behaviors using a Support Vector Machines (SVM), however, Sharma et al. [8] proposes the designing of an artificial neural network (NN) for the classification of Human activity data received from an accelerometer sensor. This kind of technique leads to good performance, but the results are not easily interpretable [9]. Another suitable choice is a probabilistic approach. An illustration is given by Van Kasteren [10], in which Dynamic Bayesian networks are used to recognise activities. In particular, temporal probabilistic models have been shown to give a good performance in recognizing activities from sensor data, as shown in Patterson's work [11]. Because of the intrinsic nature of activities, a hybrid approach is often used, which combined boosting and learning an ensemble of static classifiers with Hidden Markov models (HMMs) to capture the temporal regularities and smoothness of activities. Moreover, in these two different approaches, it is difficult to take into account a previous high-level knowledge. This aspect could be easily introduced in a logical approach using an ontology, but the main lack of logical methods is the difficulty to manage uncertainty. To overcome any limit, the three approaches have been combined, by Getoor and Taskar [12], in Statistical Relation Learning (SRL) that integrates elements of logic and probabilistic models.

In our work, as introduced before, we propose a methodology to discover human activity that integrates ontology with CEP engine based on data stored in a database produced by a wide range of sensors. So, the proposed approach includes different areas of research: domain ontology, mapping database to ontology, semantic data pre-processing, pattern matching and discovery in data words.

Ontologies are commonly used to explicitly formalize and specify a domain of knowledge [13]. Furthermore, they improve the automation of integration of heterogeneous data sources, providing a formal specification of the vocabulary of concepts and their relationships as described in the Gagnon's work [14]. A wide literature on the use of ontologies for information integration over various domains is available. In particular, an ontology for smart home is defined by Bonino et al. [15] for formally expressing what they call the "domotic environment" (e.g., sensors, gateways and network), but also for supporting reasoning mechanisms. The reasoner allows to support automatic recognition of device instances and to verify the formal correctness of the model.

Other interesting works presenting ontologies for AAL activities are those by Mocholi et al. [16] and Gu et al. [17]. More specifically, in the last work the authors present ontology-based context model using OWL, which ontology is divided into upper and domain-specific ontologies. The former is a high-level ontology and it is able to describe general context knowledge about the physical world, whereas the latter defines the details of general concepts and their properties in each subdomain. Instead, for mapping an external database to a local ontology, we refer to techniques suggested by Sedighi [18] and Barrasa et al. [19]. In addiction, tools that automatically generate OWL ontologies [20] from database schemas have been also presented, for instance by Cullot et al. [21] and Rodriguez-Muro et al. [22]. Furthermore, ontologies may also support a semantic approach to applications involving Business Process Management (BPM) techniques and analyses of processes based on a list of recorded events, i.e., Process Mining. In this case, a possible procedure is to enrich the event logs coming from external data sources by using ontology based data integration, as observed by Tran Thi Kim and Werthner [23]. A similar methodology used to integrate semantic annotation to the event log is illustrated in a BPM

context by Ferreira and Thom [24], where semantic reasoning is used to automatically discover patterns from the recorded data.
In the field of activity recognition, time interval restrictions become essential. Cases of dealing with complex events are rapidly increasing. To address this issue, ontologies are used as a basis to preserve information and relationships among events. Thereafter, they are temporally managed by a Complex Event Processor (CEP), yielding to a semantic complex event processing technique as proposed by Taylor and Leidinger [25], where ontologies are used used only for event definition and CEP tool for stream processing.

III. METHODOLOGY

In this section, we will describe our methodology, together with motivitation that drives its development. Furthermore, after the description of the architecture of the framework, we will expound the components of the system in details.

A. Motivation

The main goal of AALISABETH project is to detect a set of abnormal behaviours that could bring to the onset of most common diseases. To reach this aim, a technology based on the use of non-invasive sensor network was developed. In particular, recognise human activities by monitoring which home appliance is in use and how long user spends time on appliances is our goal. So, in the private environment a wide network of smart objects is installed. More specifically, a set of sensors has been selected and interconnected through an heterogeneous communicating network that wires the AALISABETH Smart Home (SH). Then, the wide amount of heterogeneous data generated from such network is stored in a data repository, appropriately developed as described in the following sections. They show a fine-grained nature, carrying generally their value, the originating device, the data type, the timestamp and so on. However, the acquired data show a lack of information, both the semantics and the relationships among the inert and living objects that furnish the smart home. Furthermore, one should focus not only on the single values of data, but rather on its meaning within the context. In order to take into account such relationships and formalize the knowledge of the whole context, we propose a methodology based on the employment of a specific domain ontology and it makes use of a CEP engine. In particular, the ontological modeling allows to explicitly specify the key concepts and their properties for a given domain, initially the resulting ontologies are essentially knowledge models. Furthermore, data generated by real domain can be loaded in this model, so an ontology model allows us to merge both static and dynamic information. The granularity features of acquired data are a stumbling block for the contained semantic information, which may be eventually lost. Also, a further verifiable aspect is data redundancy; that is, there can be several devices, which apparently output different results, but they provide the same information. Hence, the ontology is introduced to somehow circumvent such technical aspects and to form a bridge from the real-world system and its formal representation. In fact, it is able to merge the static knowledge and the dynamic parts by means of classes and their instances, rebuilding the whole context. Therefore, the advantages of a semantic technique are exploited twice. Once the ontology-based method has provided a conceptualization and specific description of the real-world system, such formalization drives the analysis phase. In our specific case, it is needed to look for well-determined set of data. It is worth noting that such research has to be performed according to the own semantics of the desired set. This requirement represents the main reason why an ontology-based technique is introduced.

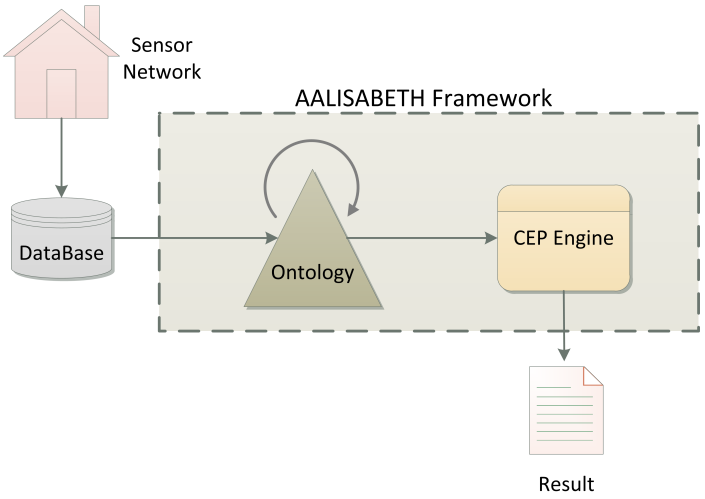


Figure 1. A simplified Architecture Model.

B. Architecture of the Framework

In order to address the situation previously described, we propose the framework depicted in Figure 1, in which main components of the architecture model are shown. As it is possible to note, the framework is essentially composed of two different interconnected components: Ontology and CEP Engine. More in detail, first of all, data collected in a MySQL relational database management system (RDBMS) are mapped in OntoAALISABETH, a domain ontology for AALISABETH project. In order to create a correspondence from records of DB to individuals of ontology, a d2rq language is used [26]. In particular, it is able to support conditional mappings, mapping of multiple columns to the same property, the handling of highly normalized table structures where instance data is spread over multiple tables, and the usage of translation tables in the mapping process. In this way, a correspondence between each element of DB and the ones of the previously implemented ontology is established. Thanks to the features of the ontology, data can be achieved efficient interoperability of information and they can be reorganized according to their semantic meaning because during the mapping phase it is possible to define how extract data. Moreover, by deploying the rules capabilities, we can divide and combine records. In fact, it is possible to build specific semantic rules in order to organize records.
The next task is represented by the simplification and aggregation of data so the fine-grained nature of data stored in the DB become a list of events that are provided by "virtual sensors". A virtual sensor represents a fictitious sensor, whose data are an appropriate aggregation of sensors data (from one or more different sensor type). For example, in order to detect the "toilet usage" action it is necessary to find a set of atomic actions and a set of locations to identify each particular activity. Furthermore, such events are occurred in specific time. More specifically, it is necessary to getting up, going to bathroom, using the

record_id	record_timestamp	host_id	obj_id	var_id	user_id	timestamp	data	int_value	real_value
199548	2015-02-02 01:28:58	1005	10011	20013	NULL	2015-02-02 01:28:58	2	NULL	NULL
199552	2015-02-02 01:29:07	1005	10011	20013	NULL	2015-02-02 01:29:07	2	NULL	NULL
199553	2015-02-02 01:29:12	1	107	17	2	2015-02-02 01:29:12	NULL	-51	NULL

Figure 2. A partial DB view.

toilet flush. So, the virtual sensor "toilet usage" collects data pertaining to involved sensor for such activity. The generation of these events is critical because, as described later, the pattern recognition is made by a specific event processor. In fact as far as time constraints are not taken into account, an ontology is sufficient to classify and organize data produced from both physical and virtual sensors. Moreover, it is able to achieve efficient interoperability of information systems. However, since our final aim is to obtain a specific time-dependent output, we need to introduce in our framework a component able to manage these time restrictions. This issue is solved by the use of a Complex Event Processing (CEP) engine, that is, a technique concerned with timely detection of compound events within streams of simple events [27]. In wider terms, the scope of this engine is to identify meaningful events.
Now we are entering into a more detailed description over every single component.

C. Components of the AALISABETH Framework

1) Database: The core of the AALISABETH system architecture is represented by the very heterogeneous sensor network, whose data produced are stored into a classical SQL Database, as introduced in Section I. Such database is developed in order to allow the integration between the different elements of embedded network and the supervision system. Moreover, its structure is organized to facilitate the extraction and interaction of contained information. In addiction, due to heterogeneous of devices, it is necessary that the database contains also information about characteristics of each device. So, it is made up a set of table, each of them contains particular information. As described above, data are collected by the sensor network into a classical SQL Database. The most important table is, clearly, the data table that contains all the records. In Figure 2 it is possible to see a small portion of that table. There are various useful columns, the following is a brief description:

- *Record_id*: the unique record id;
- *Record_timestamp*: the time of writing into the DB;
- *Host_id*: the host gateway that writes the record;
- *Obj_id*: the sensor device id;
- *Var_id*: the variable id of the measure;
- *User_id*: the user id acquired by sensors;
- *Timestamp*: the time of acquiring values;
- *Data*: a string column containing sensor measure;
- *Int_value*: an integer value of the measure;
- *Real_value*: a double value of the measure.

In the list of columns, it is possible to note that there are three different type of measure (a string, an integer, a double) this is due to the different nature of sensor devices and the physical value to measure. Also for this reason we have introduced the ontology, in order to filter and standardize data.

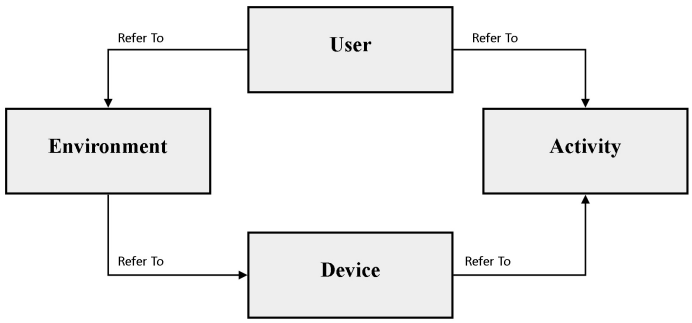


Figure 3. Context ontology overview.

a) Ontology Structure: In our proposed framework, the main element is represented by the ontology that clearly defines the semantics of the considered domain and is used as a shared knowledge base for all related components.
Moreover, it has the ability to reuse knowledge and integrate several knowledge domains. On the other hand, the AAL system is very open and it is able to change. For these reasons, a specific domain ontology, called OntoAALISABETH, has been developed. It shows a particular structure as illustrated in Figure 3, in order to model the whole system. The core is represented by the user that performs activities in own home, monitored by a sensor network embedded in the environment. In accord with this observation, the ontology is composed by four main domain components connected each other. More specifically, User, Environment, Activity and Device are the main parts of AAL system. In fact, User describes the concepts related to user's profile, e.g., age, weight. Another important information can be represented by the medicines that user needs to take, number of children or the presence of a pet. In summary, the knowledge contained in the subdomain is related to user's profile and subsequently it is connected with the performed activities. Each activity is identified by a set of atomic actions and locations. Moreover, it is performed in a time period and it is characterized by a duration. So, in the ontology, it is described in terms of locations, atomic actions, time period and duration.

These two parts of the ontology play the central role. Consequently, the appliances within the AAL environment should adapt to the user, and not vice versa. Then, Environment and Device describe user's house and the sensor network installed. Furthermore, this ontology shows different abstraction layers that composed together form a pyramid-like structure, where each lower level specialises the one on the next upper layer.

The architecture, as reported in Figure 4, is realized by the

following main components:

- A static layer (domain and domain-specific ontology);
- A dynamic layer (data and view ontology).

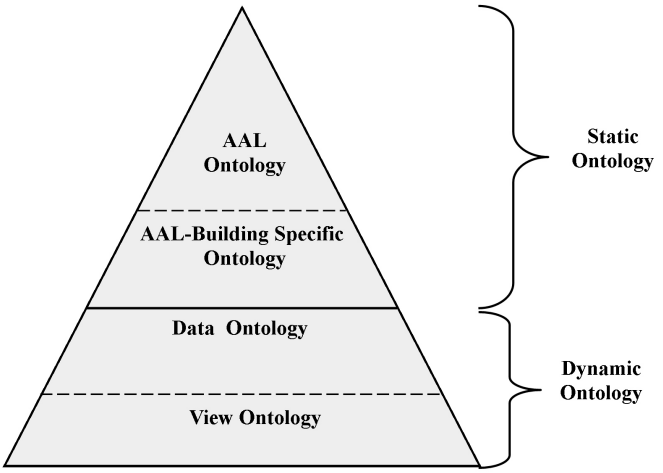


Figure 4. Pyramid-like structure of the ontology.

Each part of our ontology plays a specific role in order to respond to different requirements of the project, as described below.

b) Domain ontology: Initially, an upper domain ontology is built. One should note that this higher level of abstraction can be considered as a ready-to-use ontology for any other analogue domain. In other words, it consists of an ontology, which generally formalizes concepts present in some context, and is thought to be commonly valid. In fact, concepts are described as much generally as possible, carrying static information. Since our instance is an AAL context, as the literature suggests, we implemented a domain ontology extending and reusing an existing one. In our case, the starting ontology to model environment has been chosen to be DogOnt [15]. It has been built in a smart home context, but does not take into account several elements of an AAL environment. Therefore, we have formalized classes and relationships about the SH, its architecture and furniture, the presence and activities of one or more users, the introduction of smart objects with a communication network, sensors and clinical devices, and so on. Specifically, one imported from DogOnt ontology and it is composed of:

- **Building Thing**, it describes all the elements of a Building Environment, divided into Controllable and Uncontrollable elements;
- **Building Environment**, it models rooms and architectural spaces that compose an house;
- **Functionality**, it shapes the ability of a device to be controlled and it defines the possible commands and their range;
- **State**, it classifies continuous or discrete states, according to the kind of values they can assume;
- **Notification**, it models the ability of a device to issue a notification about state/configuration changes and it defines the corresponding notification.

The other category that has been introduced to fully describe an AAL context is defined by the following classes:

- **Activity**, it models all main daily human activities: sleeping, preparing and having a meal, walking, etc;
- **Consumable Thing**, it describes the main categories of foods, drinks and medicines;
- **Environment Profile**, it is divided in two classes: Person and Natural. The class Person depicts the users' main characteristics (as weight, age, build, etc) instead the class Natural is divided in two subclasses: Season and Weather, with the aim to take into account external environmental conditions;
- **Meal**, it introduces the different repast during a day, as breakfast, lunch, etc.

c) Domain-specific ontology: This first middle layer places below the previous upper ontology, extends several static properties and focuses on the structure of the considered domain. In our domain-specific ontology, we formalize the various components belonging to the home environment: the real structure of the ambient and disposition of rooms, the personal information about who lives in the house, which sensors are installed in the network and how they communicate. Also, the complete knowledge of the domain allows the developer to add new elements and relationships in the ontology, which cannot be described in the technology of data storing.

d) Data ontology: The data ontology extends the previous domain-specific layer introducing the concept that each device generates fine-grained data. In this level, the described classes are instantiated with individuals that present a one-to one correspondence with each record stored in the DB. This procedure is obtained via the use of d2rq language. It consists of a mapping that associates data from data sources with concepts in the ontology. Hence, the whole data ontology is implemented taking into account the sensor network, formalized in the previous layer, and is continuously updated. In this step, the semantic information about the fine-grained data is partially recovered, but the following layer permits to have custom specific views of the system.

e) View ontology: In our system, data are generated by the pervasive network, which is installed to monitor user lifestyle. In particular, such records may assume different meanings depending on the specific context. For instance, if a presence in the bedroom is followed by one in the kitchen, it has a different meaning from the same followed by one in the bathroom. Since a particular record deserves different semantic treatments, the view ontology takes into account such various circumstances. More frequently, one must evaluate the presence in the bedroom from different points of view. In terms of an ontology, this necessity converts to the implementation of new view classes where individuals are inferred. So, alternative views provided by this lower layer are needed in order to reorganize instances of data ontology. These views are defined by the expression of several equivalent classes. They are driven by the main scope to classify instances having well-determined properties and relationships; that is, these classes are populated by the desired individuals and carry the same knowledge replicated several times. The whole process of reorganization is allowed by the use of the reasoning tools, which represents the formal basis for the expressive strength of OWL. In fact, through this instrument, it is possible to obtain

additional statements that are inferred from the facts and axioms previously asserted. This reviewing step is the grounding of the preprocessing procedure. Thereafter, the reasoning tool allows to perform semantic queries on the ontology and extract the desired information for the following effective analysis, as reported in Figure 1. One should note that querying the ontology in this final step of the proposed methodology corresponds to select an amount of data generated by virtual sensors, i.e., a group of data following the user interpretation of the system. Moreover, this approach developed by means of inference classes has the important advantage to be extensible and additive. In order to better explain the advantages deriving from the classification of the view ontology, let us consider the following cases. One of the most relevant aspects of our project is the capability of monitoring if the user gets up during the night for eating or toileting. In order to recognize these activities, we proceed creating two views, i.e., macro ontology classes. Each class contains all inferred individuals that allow the eventual recognition of the considered activity. In this particular case, the information about getting up and exiting from the bedroom are common. Instead, presence and utilization of the toilet is found in the first case, while presence in the kitchen and opening a sideboard or refrigerator belong to the second view. Furthermore, in both cases we require that the person comes back to the bedroom after some time and continues to sleep. Hence, these sets of individuals populating the view classes are selected as input for the following step of analysis. It is worth noting that processing data with the described technique allows to preserve relationships and constraints introduced by the previous domain-specific layers of the ontology. Contents of each layer of the pyramid-like structure are shown in Figure 5.

D. Framework Implementation

Let explain more deeply the framework by using Figure 6, which clearly shows the dataflow of the system. The starting point is represented by a data table full of useful but also noisy data. At the real beginning, there is a simple extraction of data, every instance of the ontology represents every row of the database. The second phase is represented by filtering and selecting data in which we model data entries in a different way. In fact, the concept of "entry" is substituted by the concept of "event". By the deployment of OWL Java library, data entries are aggregated and picked out in order to generate more complex events, called "macro-event". During this step, macro-events are further selected and filtered. This could be done thanks to the deployment of ontology rules and reasoner. In fact, as described before, the importance of the ontology is its ability in preprocessing task and semantic filtering. It is possible to define rules to select, clear or divide by semantic meaning. The next task is performed by the interaction with the user, which could not be an IT expert. The user can choose two important parameters: which pattern he wants to analyze and the query string. So the system sends to the CEP engine, Esper, the events as Java POJO objects. Then Esper according to the input query string analyses the stream of events looking for patterns that match with the query.

1) Reasoning & Rules: In order to reorganize and preprocessing data, a set of rules are introduced into ontology. In particular, such rules allow to group data depending on time interval. In order to reach the aim, a set of built-in rules, an alternative paradigm for knowledge modeling, are introduced to acquire new knowledge by establishing new object property connections between unrelated entities. These rules, which are able to extend the expressivity of OWL, are evaluated periodically during runtime and new facts are added into the ontology. The built-in rules can be easily extended by defining custom rules. The definition of equivalent classes is driven by the main scope to classify instances carrying determined properties and relationships; that is, these classes are populated by the desired individuals. Instead, the preprocessing phase is based on the ability of the reasoning tool to query the ontology and extract the required information for the following effective analysis. One should note that querying the ontology in this final step of the proposed methodology corresponds to select an amount of data generated by virtual sensors, i.e., a group of data following the user interpretation of the system.

2) Complex Event Process (CEP) Analysis: The ontology structure is extremely powerful but it has serious expressiveness limitations: the lack of support of temporal reasoning. Considering the nature of our analysis and the dynamic updating of dataset, traditional methods do not allow to perform reasoning over time and space, so we introduce a CEP engine in order to perform the temporal analysis procedure. This engine permits to combine data from multiple sources to infer events or patterns that suggest more complicated circumstances. In fact, the main objective is to recognize significant events. These identifications could be eventually reused to discover further more complex events, through additional uses of CEP engine.
In our framework, we deploy Esper [28], a CEP engine library, which is able to process large volumes of incoming messages or events, regardless of whether incoming messages are historical or real-time in nature. It is successfully used into finance field (e.g., trading, risk management).

3) Java Component: The whole framework is developed using Java Language. We chose Java because using that it is possible to read and manage OWL Ontology thanks to Jena API. Furthermore, we are able to dynamically create Java classes at runtime. In fact, the initial idea is to develop a "context-free" framework, in other words the framework has not to be related to structure of the ontology, sensor type and home environment. Moreover, Esper does not inherently support a specific access to OWL ontology, so it is necessary to map the events of the ontology into simple Plain Old Java Objects (POJOs). For this reason, we implemented a component that by the deployment of Javassist library [29], is able to create Java classes corresponding to ontology classes. These Java classes are also called "POJO classes". In that way, the system results extremely dynamic regarding the variables of the environment.

4) Graphical User Interface (GUI): In Figure 8 is shown a first Graphical User Interface (GUI). It is composed of different buttons and text boxes. The execution is divided essentially in three phases:

1) by clicking the "Load Data" button, records contained into the DB are exported into the Ontology file thanks to mapping procedure;
2) on the right the user can choose which type of pattern wants to analyze;
3) then the user can define a query string on his own and wait for results after clicking the "Execute query" button.

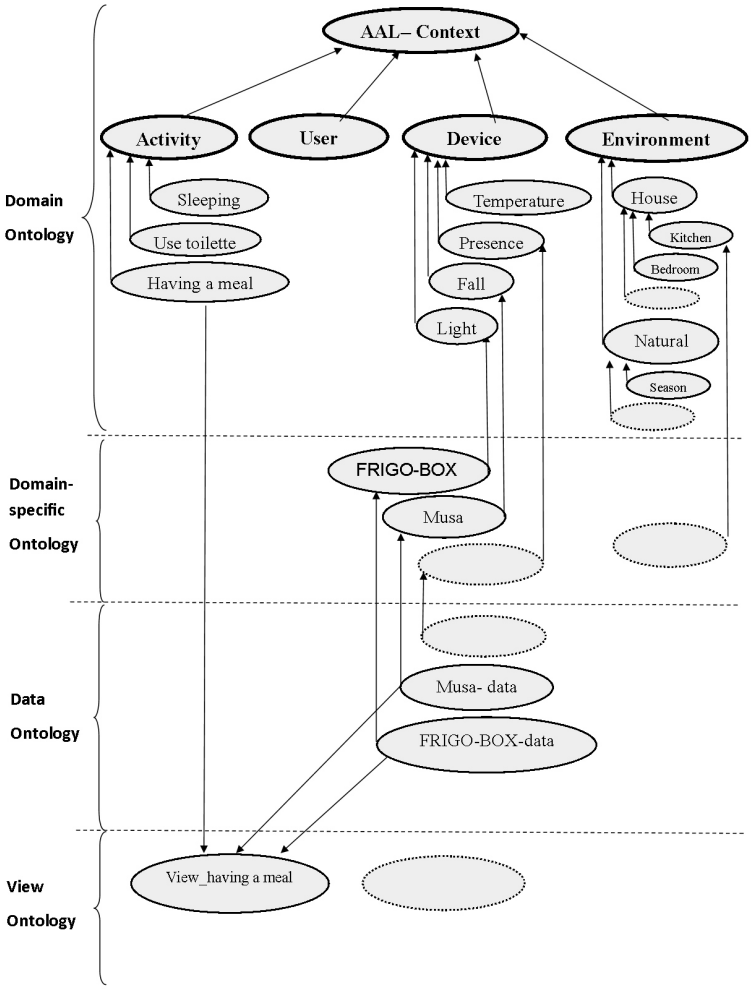


Figure 5. Class hierarchy diagram of OntoAALISABETH.

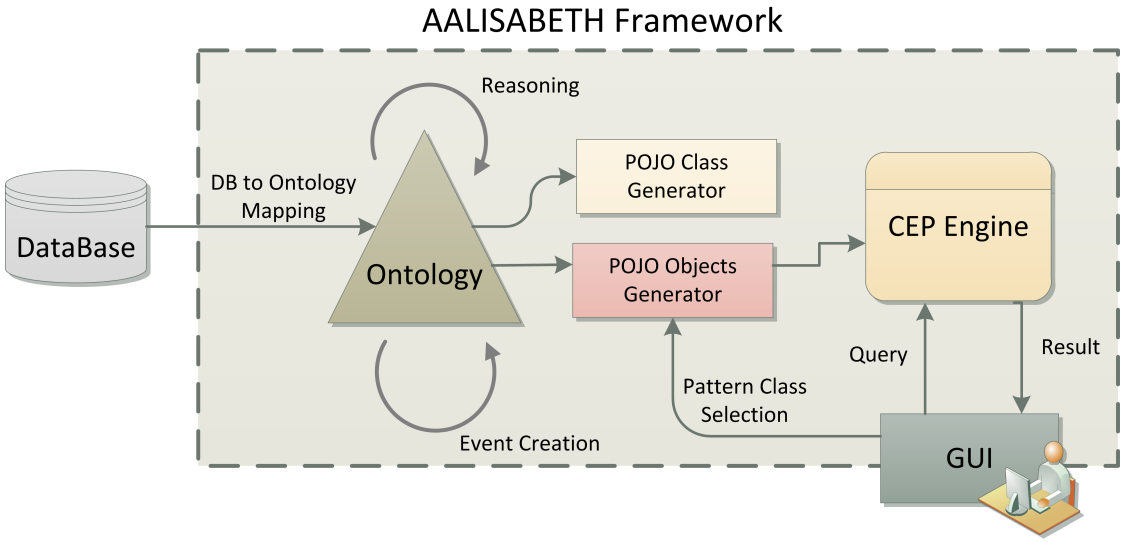


Figure 6. A detailed Architecture Model.

This is a simple GUI that allows us to deploy the main features of the framework.

5) Tools & Technologies: We are now trying to summarize tools and technologies that allowed us to develop this framework. First of all, the OWL ontology is developed and tested in Protégé 4.3 [30], together with the Pellet Reasoner Plug-in [31], which permits the creation and population of equivalent classes. Through the definition of a mapping configuration we deployed a OBDA system, in order to write down the statements that map the Database to the ontology. To implement the framework, we use Java as a coding language to combine several techniques. Then, the ontology is managed by means of the OWL API. Thereafter, the Pellet reasoner is invoked through Jena [32] to perform reasoning over the ontology together with the individuals. The SPARQL query is also executed through Jena.
Basically, using Jena we load the ontology file created with Protégé into an ontology model (a Java object implementing the OntModel interface). We then choose to utilize Esper as CEP tool for several reasons: its open-source Java library for complex event processing, it can be used in different data streams and CEP applications, it has adapters that allow the user to provide different input formats for the representation of events.
The whole Java framework is developed using Eclipse IDE [33].

E. Test & Validation

The test configuration is composed of a standard flat in which an elderly person lives. The inhabitant has a regular behaviour although he knows that a sensor network was installed and where sensors are positioned. This sensor network captures a great number of actions performed by the user. Data generated are initially store into a classical DB, then are imported into our framework.
In this section, we are going to explain, using a simple example case, capabilities and features of our approach. First, we have to describe the characteristic of the test configuration more in detail.

1) Use Case Description: Our test house is a ground floor flat, where a man lives. In Figure 7, it is possible to see the plan of the flat. It is composed of a living room / kitchen, a bedroom, a bathroom and an office.
The sensors installed are (the number in the list correspond to the number on Figure 7):

1) *Presence Couch* sensor: a sensor that measures the weight of the couch;
2) *Passage Office* sensor: observes the passage of a person coming in/out office;
3) *Passage Living (Hall)* sensor: as for sensor 2, observes the passage of a person coming in/out living room;
4) *Opening Fridge* sensor: observes when the fridge is opened;
5) *Opening Pantry* sensor: observes when the pantry is opened;
6) *Scale*: observes and captures what the user is weighing;
7) *WC flush* sensor: captures when the flushing device is used.

In this set of sensors, probably the most interesting is the Scale sensor. In fact by the use of particular plates provided by Radio-Frequency IDentification (RFID) tag, the sensor observes the weight and the type of dish. That allow us to calculate the calories and the amount of carbohydrates.
This sensor network will be our heterogeneous data source in the provided example.

2) Pattern Recognition Example: Considering the described above type of sensor available we define this query string:

```
select * from  pattern [ every (A =
    DataElementEvent(event_type=1)
    -> (B = DataElementEvent(event_type=4)
    and not DataElementEvent(event_type=1))
    -> (C = DataElementEvent(event_type=3)
    and not DataElementEvent(event_type=4))
    -> D = DataElementEvent(event_type=7
    and timestamp_start <
    A.timestamp_start + 3600000))]
```

This pattern represents a sequence of actions, starting from the bathroom, then the passage into the living room, then opening the fridge and finally being present on the couch. All of these actions must be performed within 10 minutes ($A.timestamp\_start < A.timestamp\_start + 3600000$, timestamp values are processed as milliseconds long type). In order to fully understand Esper querying syntax we remand to references [28].
In Figure 8, we want to illustrate the output of the described Esper query. Once the event set and user have been selected, POJO objects representing events are created and passed as input to Esper engine. Now, we define our query pattern and execute it. On the bottom, are displayed the results that are all patterns that match the query.
In our example, the framework was able to find two matching patterns for the input dataset, represented by five days of monitoring. As it is possible to note in Figure 8, patterns found are related to the dates "*2015-02-04*" and "*2015-02-05*".
The example exposes the main potentialities of the framework, starting from management and selection of a huge amount of data, and passing through the definition of an ad-hoc query string for every type of recognizable human behaviour. In fact, this kind of example must be considered as an initial point for a more complex and elaborated use case according to a precise and meaningful social analysis.

IV. Conclusion

In this paper, firstly we have presented an ontology-based framework to retrieve semantic information from a data repository. Later, we have illustrated a case of use in which our methodology is applied and the obtained results.
Our work represents a different approach to the pattern (or activity) recognition problem. The novelty of our architecture is represented by the combination of two concepts: semantic reasoning with temporal analysis. Due to the ability of ontology to reuse knowledge, it represents the central element of the presented methodology. The developed ontology, named OntoAALISABETH, is characterized by four layers: a top-level ontology followed by a domain-specific one, and data layer which establishes over a final basis-view layer. The top-ontology has a particular structure, in particular it is composed of four domain ontology system - User, Environment, Activity, Device - that represent the whole knowledge base in AAL domain. The last part is thought as a data preprocessing step. It plays the role to organize data according to the desired context

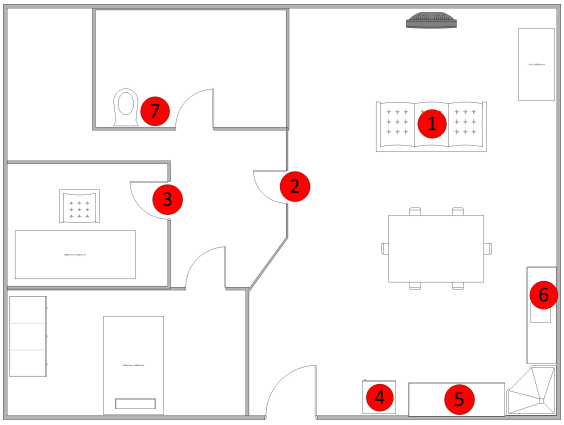


Figure 7. A plan of the flat.

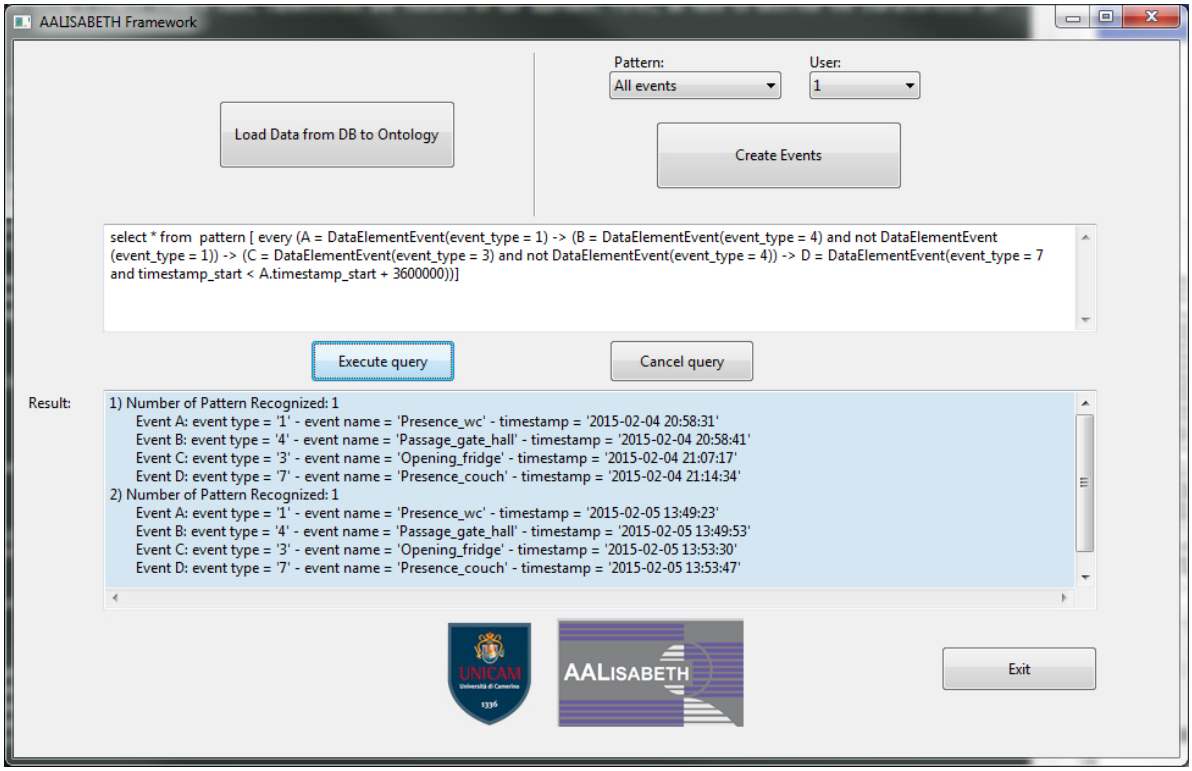


Figure 8. The AALISABETH GUI with patterns recognized.

views, in order to allow a proper analysis. Then, due to the expressiveness limitation of ontology, the Cep engine has been introduced. In fact, the CEP engine works as an analysis tool to support the timed reasoning.

In general, knowing the lifestyle of human in home is not an easy task, since each person has the different way to perform the activity. Some activity events are occurred in specific time for each day. Hence, time can be used to distinguish the activity in specific detail. For example, we can know the "Eating & drinking" activity, which is for breakfast, lunch, or dinner based on time. In addition, it is very important for further analysis, i.e., health-care system needs to know the time when patient has a meal in each day. Thus, if we know the lifestyle of human in home environment, we can predict, which activity will occur in specific time. For instance, human take a bath twice a day, after wake up and before goes to sleep. The system can predict the "take a bath" activity, if user wake up in the morning and go to the bathroom.

Unfortunately, we can not provide any comparison in term of performance with different kinds of approaches. However, the model described is very extensible and it is not bound by sensor types or environment variables.

Regarding possible future developments, it could be very interesting if the pattern found is stored into the ontology as a new macro-event with a label. In this way, it will be possible to look for particular pattern into this macro-event set. In this manner, we will create another level of abstraction with a more complex and detailed behavioural analysis. Moreover,

the main disadvantage is probably represented by the Esper query language that could be tough for an non-IT expert, for this reason, it could be extremely helpful to develop a custom pseudo-code notation, closer to human language, which will be interpreted by the framework.
A good improvement could be represented by the implementation of a real-time analysis, in other words, every time the database receives a new record, it alerts the framework that extracts the information and manages it in order to detect a matching pattern at real-time. This kind of approach will be useful also in emergency situations in which the caregiver has to rescue the user (e.g., a detected fall).
In conclusion, we have developed our work in an AAL context, but thanks to extensible nature of the framework it is reasonable to think that our approach can be applied also on smart cities, a city characterized by modern urban production factors in a common framework with the intent to improve the quality of life and a sustainable economic development.

ACKNOWLEDGMENT

The authors would like to thank every partner of AALISABETH project for the great working collaboration done until now. They also acknowledge the financial contribution of the Marche Region administration, under the action Smart Home for Active and Healthy Aging, for supporting the research on the AALISABETH project.

REFERENCES

[1] R. Culmone, M. Falcioni, and M. Quadrini, "An ontology-based framework for semantic data preprocessing aimed at human activity recognition," in The Eighth International Conference on Advances in Semantic Processing, SEMAPRO 2014. IARIA, 2014, pp. 1–6.

[2] The ambient assisted living (aal) joint programme. Last checked: 2015-05-28. [Online]. Available: http://www.aal-europe.eu/

[3] F. Castiglione, V. Diaz, A. Gaggioli, P. Liò, C. Mazzà, E. Merelli, C. G. Meskers, F. Pappalardo, and R. von Ammon, "Physio-environmental sensing and live modeling," Interactive Journal of Medical Research, vol. 2, no. 1, 2013.

[4] F. Corradini, E. Merelli, D. R. Cacciagrano, R. Culmone, L. Tesei, and L. Vito, "Activage: proactive and self-adaptive social sensor network for ageing people," ERCIM News, vol. 2011, no. 87, 2011.

[5] E. Frontoni, E. Gambi, L. Palma, L. Pernini, P. Pierleoni, D. Potena, L. Raffaeli, S. Spinsante, P. Zingaretti, D. Cacciagrano, F. Corradini, R. Culmone, F. D. Angelis, E. Merelli, B. Re, L. Rossi, A. Belli, A. D. Santis, and C. Diamantini, "Interoperability issues among smart home technological frameworks," 2014.

[6] Aalisabeth - ambient-aware lifestyle tutor, aiming at a better health. Last checked: 2015-05-28. [Online]. Available: http://www.aalisabeth.it/

[7] P. Chahuara, A. Fleury, F. Portet, and M. Vacher, "Using markov logic network for on-line activity recognition from non-visual home automation sensors." in AmI, ser. Lecture Notes in Computer Science. Springer, pp. 177–192.

[8] A. Sharma, Y.-D. Lee, and W.-Y. Chung, "High accuracy human activity monitoring using neural network," International Conference on Convergence Information Technology, 2008.

[9] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 1, Jan. 2000, pp. 4–37. [Online]. Available: http://dx.doi.org/10.1109/34.824819

[10] T. van Kasteren and B. Krose, "Bayesian activity recognition in residence for elders," in Intelligent Environments, 2007. IE 07. 3rd IET International Conference. IET, 2007.

[11] D. J. Patterson, D. Fox, H. A. Kautz, and M. Philipose, "Fine-grained activity recognition by aggregating abstract object usage." in ISWC. IEEE Computer Society, pp. 44–51.

[12] L. Getoor and B. Taskar, Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning). The MIT Press, 2007.

[13] T. Gruber. What is an ontology? [Online]. Available: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html (2009)

[14] M. Gagnon, "Ontology-based integration of data sources," in Information Fusion, 2007 10th International Conference on.

[15] D. Bonino, E. Castellina, and F. Corno, "The dog gateway: enabling ontology-based intelligent domotic environments." IEEE Trans. Consumer Electronics, no. 4, pp. 1656–1664.

[16] J. Mocholí, P. Sala, C. Fernández-Llatas, and J. Naranjo, "Ontology for modeling interaction in ambient assisted living environments," in XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010. Springer, 2010, pp. 655–658.

[17] T. Gu, X. H. Wang, H. K. Pung, and D. Q. Zhang, "An ontology-based context model in intelligent environments," in Proceedings of communication networks and distributed systems modeling and simulation conference, vol. 2004, 2004, pp. 270–275.

[18] S. M. Sedighi and R. Javidan, "Semantic query in a relational database using a local ontology construction," South African Journal of Science, vol. 108, no. 11-12, 2012, pp. 97–107.

[19] J. Barrasa Rodríguez, Ó. Corcho, and A. Gómez-Pérez, "R2o, an extensible and semantically based database-to-ontology mapping language," in In Proceedings of the 2nd Workshop on Semantic Web and Databases. Springer-Verlag, 2004.

[20] OWL 2 Web Ontology Language Document Overview, W3C Recommendation, Std., 10 2009. [Online]. Available: http://www.w3.org/TR/owl2-overview/

[21] N. Cullot, R. Ghawi, and K. Yétongnon, "Db2owl : A tool for automatic database-to-ontology mapping." in SEBD, 2007, pp. 491–494. [Online]. Available: http://dblp.uni-trier.de/db/conf/sebd/sebd2007.html#CullotGY07

[22] M. Rodriguez-muro, L. Lubyte, and D. Calvanese, "Realizing ontology based data access: A plug-in for protégé," in In Proc. of the Workshop on Information Integration Methods, Architectures, and Systems (IIMAS 2008. IEEE Computer Society Press, 2008, pp. 286–289.

[23] T. Tran Thi Kim and H. Werthner, "An ontology based framework for enriching event log data," in SEMAPRO 2011, The Fifth International Conference on Advances in Semantic Processing, 2011, pp. 110–115.

[24] D. R. Ferreira and L. H. Thom, "A semantic approach to the discovery of workflow activity patterns in event logs," International Journal of Business Process Integration and Management, vol. 6, no. 1, 2012, pp. 4–17.

[25] K. Taylor and L. Leidinger, "Ontology-driven complex event processing in heterogeneous sensor networks," in Proceedings of the 8th Extended Semantic Web Conference on The Semanic Web: Research and Applications - Volume Part II, ser. ESWC'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 285–299. [Online]. Available: http://dl.acm.org/citation.cfm?id=2017936.2017959

[26] The d2rq platform. Last checked: 2015-05-28. [Online]. Available: d2rq.org/

[27] D. Anicic, P. Fodor, S. Rudolph, and N. Stojanovic, "Ep-sparql: a unified language for event processing and stream reasoning," in Proceedings of the 20th international conference on World wide web. ACM, 2011, pp. 635–644.

[28] Esper 5.2 documentation. Last checked: 2015-05-28. [Online]. Available: http://www.espertech.com/esper/documentation.php

[29] S. Chiba, "Javassist - a reflection-based programming wizard for java," in International Business Machines Corp, 1998.

[30] Protégé - a free, open-source ontology editor. Last checked: 2015-05-28. [Online]. Available: http://protege.stanford.edu/

[31] Pellet reasoner plug-in for protégé 4. Last checked: 2015-05-28. [Online]. Available: http://clarkparsia.com/pellet/protege/

[32] Apache jena - a free and open source java framework for building semantic web. Last checked: 2015-05-28. [Online]. Available: http://jena.sourceforge.net/

[33] Eclipse ide. Last checked: 2015-05-28. [Online]. Available: https://www.eclipse.org/

Detection of Floor Level Obstacles and Their Influence on Gait

A Further Step to an Automated Housing Enabling Assessment

Nils Volkening
Department of Health Services Research
Carl von Ossietzky University Oldenburg
Oldenburg (Oldb.), Germany
Nils.Volkening@uni-oldenburg.de

Andreas Hein
Department of Health Services Research
Carl von Ossietzky University Oldenburg
Oldenburg (Oldb.), Germany
Andreas.Hein@uni-oldenburg.de

***Abstract* - The demographic change in the industrial countries is a great social challenge. To ensure constant or better (health) care in the next decades, new care concepts for older people are needed. An approach is the use of Information and Communication Technology based solutions. Especially the preservation of personal mobility should be in focus because it is a key role to sustain autonomy and social interaction of senior citizens. In addition to the age-based declining mobility, there are secondary events, which reduce the mobility of senior citizens, e.g. diseases or fall events. Prevention of fall events is a goal for the Housing Enabling Assessments by adaption of room, e.g., by detecting and removing tripping hazards. Former work proves that an automated Housing Enabling Assessment executed by an autonomous service robot could achieve better quality and higher acceptance than a manually controlled Housing Enabling Assessment. In this article, two different methods for detecting relevant unevenness of floor in home environments and resulting challenges are presented. An adapted autonomous service robot is used as well as a Microsoft® Kinect for gait analysis and, regarding the detection of the floor's unevenness, a Prime Sense ® Carmine 1.08 depth sensor and a self- designed triangulation laser scanner were compared. First results indicate that floor characteristics have a relevant influence on gait parameters, such as gait speed, step / stride length and their variation. Also, results show that floor characteristics should become a mandatory factor for in-home gait analysis.**

Keywords-mobile robot; gait analysis; floor level; RGB-D camera; triangulation laser scanner.

I. INTRODUCTION

This article is based on the AMBIENT 2014 conference paper [1] and provides an extended approach to detect floor level obstacles and further results to their influence on gait.

Industrial countries face different challenges caused by the demographic change [2]. A possible way to cope with these upcoming problems is the use of Information and Communication Technology (ICT) in the area of Ambient Assisted Living (AAL). There are two main approaches to bring technology to the homes of senior citizens. The first approach is Smart Homes [3], which means that the entire technology is integrated into the apartment. The second solution would be autonomous service robots. In this case, sensors, actuators and computational units are attached to a mobile base. An example for "simple" household robots are autonomous vacuum cleaners. Because of a great sales volume of autonomous vacuum cleaners in the last years, they have a big impact on society. They have raised the acceptance for robots among users and show how the design influences it [4][5][6]. Advanced systems like service robots could support caregivers to help elderly maintain an independent lifestyle and preserve their indoor mobility up to a high age [7][8]. A potential advantage of service robots compared to Smart Homes is their low costs since they need fewer sensors to generate a good coverage based on their mobility. In order to cover areas, they can bring them in the area of interest [9]. In this approach, the mobility of these platforms is used to realize an automated Housing Enabling (HE) assessment [10]. A first step is the evaluation of the apartment, especially the examination of the floor in order to detect stumbling risks. This article is organized as follows; Section II motivates the topic and is followed by the State of the Art and current limitations (Section III). In Section IV, two approaches are presented to measure the unevenness of floors and the measurement of different gait parameters followed by the results in Section V. The conclusions and further steps complete the article (Section VI).

II. MEDICAL MOTIVATION

Prevention of fall events is an important research area. A fall event could have great impact on mobility, especially for senior citizens. An obvious fact is the reduction of mobility in case of a fracture of the neck of femur. But also the fear of, e.g., a second fall limits the mobility of older people [11] and a reduced mobility increases the risk of falling, which is the starting point of a vicious circle. Also, fall-related costs are a major factor for our health care system [12].

An important factor is that mobility problems reduce the personal radius of movement. Renteln-Kruse et al. show that this influences social participation; above the age of 55 years, the radius of movement is reduced to approximately 3 km around the home [13]. Also, 55% of fall injuries occurred inside the house [14], which raises the importance of in-home assessments. From a clinical perspective, long-term monitoring of changes in mobility has a high potential for early diagnosis of various diseases and for the assessment of fall risk [8]. As important as the age and potential diseases / disabilities of the patient [15][16] is the condition of the floor

for the self-selected gait velocity and, in general, the risk of stumbling or slipping [17]. Especially in an unsupervised environment, the additional information about the quality of the floor could increase the precision of the gait analysis [18][19], which could be very helpful for the HE Assessment in order to estimate the personal factor. This approach, tries to realize both, i.e., a good data base for the HE Assessment and also gain additional information for a gait analysis to increase their precision.

III. STATE OF THE ART

This section gives an overview of the four most interrelated research areas of HE Assessment. First, the trend analysis of mobility in domestic environments is outlined, followed by mobility assessments using mobile robots. Afterwards, possible environmental hazards and housing modification are shown. Fourth point is the influence of the unevenness of the floor on gait parameters. Finally, the section is closed by the current limitations of the State of the Art.

A. Trend Analysis of Mobility in Domestic Environments

Various approaches for gait analysis in domestic environments are presented. Scanaill et al. present the possibility of upgrading a home with various sensors, especially from the home automation or security domain to a (health) Smart Home [20]. Most systems are used for trend analyses [21][22][23] and only some approaches use ambient sensors for detailed gait analyses [24]. Various groups use Home Automation Technologies like motion sensors, light barriers or reed contacts placed in door frames or on the ceiling. Cameron et al. use optical and ultrasonic sensors [21], which were placed on both sides above the door frames to determine the walking speed and direction of a person passing. Kaye et al. presented an intelligent system for assessing aging changes [22]. For the study, they installed several sensors in 265 homes for an average of 33 months and used, among others, wireless passive infrared motion sensors, which were covering different rooms of an apartment. A line of these sensors was modified and attached to the celling of some rooms within the apartment to estimate the resident's walking speed. Also, laser range scanners are used for different assessments. Frenken et al. presented an automated Time Up and Go (TUG) Assessment. Therefore, the laser range scanner is mounted underneath a chair and is used to recognize the legs of the test person [24]. Pallejà et al. have a similar approach but conducted a detailed gait analysis with a laser range scanner, which was mounted at 100 mm above the floor [25]. This low position of the laser has the disadvantage that it is possible that a foot could hide the corresponding leg. In this case, a laser scanner would only detect the tip of the foot and not the leg, which is important for a correct assessment. Poland et al. used a camera attached to the ceiling, recording a marked floor evenly divided into rectangles to estimate the gait speed [23]. Each of these rectangles is defined as a virtual sensor. For persons within these, the approach 'activates' the virtual sensor in, which they are currently located in. Stone and Skubic used the Kinect to analyze the gait in a home environment [26]. Especially the variation of gait parameters like step length and self-selected speed over time were measured and identified as independent factors for the personal stumbling risk. Also, Gabel et al. used the MS Kinect for a full body gait analysis, which is capable of a precise in home gait analysis [27]. A similar approach for a long-term in-house gait analysis by using the Kinect was published by Stone and Skubic [28]. But in addition, a monthly fall risk assessment protocol was conducted for each resident by a clinician, which included traditional tools such as the Timed Up and Go and habitual gait speed tests. Afterwards they compared the results of the clinician with their approach.

B. Mobility Assessments Using Service Robotics

Service robots combine ideas of different fields of robotic research into one system in order to target a specific application. Most available platforms are still in (advanced) research states. There are different fields of interest, e.g., acting autonomously in home environments. For most mobile robot platforms it is difficult to interact with the human friendly environment. A closed door could be a problem for a robot. Petrovskaya and Ng present a probabilistic approach on how a mobile robot could detect and open doors [29]. Also, the interaction with humans is very important; Breazeal published a first approach on how to design a sociable robot and how it can learn from environmental factors and user behavior [30], this approach is similar to Ray et al. [31], who asked "What do people expect from robots?". To be able to interact with humans, it is very important for the robot to be able to recognize humans. Udsatid et al. present an approach of a mobile robot platform, which tracks humans and is able to drive side-by-side with a human by using a down facing Microsoft Kinect sensor to track the feet, to find out the heading and direction of the human [32]. Brell et al. presented a first approach of a mobility assessment with a mobile robot platform [33]. For this, a laser range scanner to detect the residents 'legs and to estimate the walking speed within different areas of the apartment was used. Within our own work, a new approach on how to enhance mobile robot navigation in domestic environments by use of mobility assessment data was recently presented [34]. The advantage of a mobile robot is that it can bring the needed sensor technology to the Optimal Observation sLots (OOL) for monitoring as introduced in [33]. In the observation phase, the robot stands at a safe place in the initial room of the apartment and observes the human behavior and environment. These data are used to compute new OOL, which fulfill different safety and quality criteria. After that phase, the robot will travel to the respective OOL and measure different gait parameters by using the laser range scanner and the Kinect, which can be used in HE Assessment.

C. Environmental Hazards and Housing Modification

T. M. Gill et al. presented a study, which sought to estimate the population-based prevalence of environmental hazards in the home of older persons [35]. Therefore, one thousand homes of senior citizens above 72 years were

assessed. The most potential hazards are slip and trip hazards by rugs, carpets, etc. In second place are blocked pathways by e.g., small objects or cords and in third position insufficient lightning conditions (shadows or glare), curled carpet edges or other tripping hazards. T. M. Gill et al. pointed out that safety awareness at home may relate to one's personal capabilities. On the other side, M. E. Northridge presented a study on home hazards and the role of health and functional status of senior citizens [36]. It was pointed out that the presence of home hazards influence vigorous elderly persons twice in aspect of falling but it was not associated with the increased likelihood of falls among frail older persons. A quite popular assessment in the Scandinavian countries is the HE Assessment. It reduces the risk of fall in home environments and the near surrounding. The apartments are assessed depending on the personal health status of the residents and the structure of the apartment itself [37]. This rating gives advice on how to change the apartment with its furniture etc. so that it is suitable for the resident. The HE Assessment is split into three parts. The first part is the descriptive part for collecting general information on the apartment and the resident's condition. The second part is the evaluation of functional limitations and dependence on mobility aids. Also, detailed information about the medical condition of the user is collected, e.g., severe loss of sight or limitation of physical fitness. The last part is based on different questionnaires, which relate to the apartment and the vicinity. After completion of all questions, a score of the apartment in relation to the actual health status of the resident [38] is computed [39]. A customization of the apartment to reduce the risk of falling is also possible. This adaption is related to the rating [40] but is not an explicit part of an HE Assessment. Another survey to investigate the prevalence of environmental hazards in the homes of older people was presented by S. E. Carter [41]. This survey shows that 80% of the 425 inspected homes had at least one, and nearly 39% had more than 5 tripping hazards, while 62% showed "flooring" hazards. R. Cham and M. S. Redfern measured the change in gait when people anticipated slippery floors [42]. Therefore, three different floorings were used with the participants having to walk over each surface three times. In the first trial, the test person knew the floor was dry, next the test person was uncertain about the floor's condition (dry, wet, oily, soapy) and in the last try the condition was also known as dry. They found significant changes in the normal stride length and stance duration. It was also pointed out that the floor type had some influence on most gait variables.

D. Evenness of the Floor and its Influence

Also, the unevenness of the floor influences on gait parameters. S. B. Thies et al. reported on the effects of surface irregularity and lighting on step variability during gait [18]. Different gait parameters from 12 healthy young women and 12 healthy older women were measured. Each person had to walk over a 10m walkway in a personal comfortable speed with four different settings being tested: plain surface with regular lighting; plain surface with low lighting; irregular surface with regular lighting; and irregular surface with low lighting. As a final result, the lighting did not have a significant effect on any of the gait parameters, while the surface type had significant effects on the step time variability, step width variability, which was observed especially with the older women. Marigold and Patla also presented results on age-related changes in gait on multi-surface terrain [19] using a more outdoor-based scenario so that the multi-surface terrain consisted of solid, flexible, rocky, irregular, slippery, and uneven surfaces. Ten younger and ten older adults were tested and it was found that the step length, trunk pitch and roll, and head acceleration variability were increased on the multi-surface terrain compared to solid ground trails for both young and older adults. Older adults obtain a larger medial-lateral trunk center of mass acceleration Root-Mean-Square (RMS) and trunk roll RMS when walking on the multi-surface terrain. But they found no age-related differences in the step variability. The influence of an irregular surface and low light on the step variability of patients with peripheral neuropathy was researched by Thies et al. [43]. Also, the change in gait parameters by stepping over an obstacle was presented by different research groups using obstacles with a height between 0 mm and 152 mm [44][45]. McFadyen and Prince used an 11.75 cm height obstacle [46]. All studies measured differences in the gait patterns in general but they do not have a common result. The influence of surface slope on human gait characteristics was presented by Sun et al. [47]; for this study an outdoor set-up was used, so that the results are not exactly comparable to indoor set-ups. Nevertheless, all studies have shown that the surface does have an influence on gait. In order to estimate a maximum permissible value of the unevenness in homes, several building regulations are inquired [48][49][50][51]. They identified different levels of unevenness, which should not be exceeded. In general, all building regulations pointed out that office floors do not have uneven areas, no slots, stumbling areas or dangerous slopes. The maximum height difference between two even rooms are defined as 4 mm within the "Professional association rules for safety and health at work" [48] and the "Technical Regulations for Workplaces" [50]. The unevenness of the floor is only named as an environmental risk but has not exactly been defined. Also, the "Slip, Trip, and Fall Prevention" Guide from the University of Stanford [49] named some trip and fall hazards, e.g., uneven walking surfaces, holes, changes in level, broken or loose floor tiles, defective or wrinkled carpet or uneven steps/thresholds but it also has no exact dimensions for the different points. The DIN 18202 gives precise dimensions for evenness of the

TABLE I. LIMITS FOR FLATNESS TOLERANCES

Description	Limit of unevenness in mm among measurement distance in m				
	0.1	***1***	***4***	***10***	***14***
Screeds to receive e.g., floor coverings, flooring, tiling	2	4	10	12	15
Finished grounds with increased requirements	1	3	9	12	15

a. Excerpt from the DIN 18202:2013-4

floor (see Table I) [51]. If you measured a distance of 1m, a level difference of about 4 mm is tolerable for normal screed. These values are used as a reference point for the accuracy of our approaches. As shown before, it is important to have detailed information on the surface (floor) to raise the validity of domestic gait analysis. Udsatid et al. used a mobile robot and a Kinect sensor to measure the ground and calculate a virtual ground plane [32]. But, only for a background subtraction for a foot tracking algorithm, which was used for a side-by-side navigation algorithm. Currently, there are no mobile service robots to determine the unevenness of the floor.

E. Limitation of the State of the Art

As shown in Section III-A, most of the systems used ambient sensors and did not observe the user continuously but only measured the test person's presence at specific points. The problem herein is that it can only be used for trend analysis rather than for a detailed assessment to determine various mobility parameters of a person. For precise assessments of mobility, laboratory equipment and well-known vicinities are needed. On the one hand, the installation efforts and costs are too high to install such in domestic homes, on the other hand homes are dynamic, this means that, e.g., furnishing changes over time. All of the automated gait analyses do not respect the influence of the floor cover. Within the domain of health care and rehabilitation service robotics, there are quite few systems commercially available. Moreover, there is no robotic system, which is capable of performing HE Assessments and giving advice on how to reduce the risk of falling. The current HE tests suffer from some drawbacks, e.g., the estimation of personal disorders, the investigation and also the following customization of the apartment, which highly depends on the skill of the person executing the test. Little knowledge could lead to different or insufficient results. Furthermore, these assessments are mostly not done as a continuous assessment but rather as an event-triggered assessment after accident. In summary, there is currently no system or approach available, which is capable of conducting precise and continuous HE Assessments in domestic environments and using this additional information to raise the precision of gait assessment results.

IV. APPROACH

We are going to present two different approaches to detect relevant unevenness of the floor with a mobile robot platform. This is followed by a short description of how to estimate balance parameter under different environment conditions.

A. Detection of Unevenness with RGB-D Camera

Our first approach provided an automated and continuous detection of relevant unevenness of the floor assembly, which will be used to rate the apartment during the HE Assessment and to increase the quality of the gait analysis. In order to implement a stable algorithm in an unsupervised environment, an initial self-calibration was included.

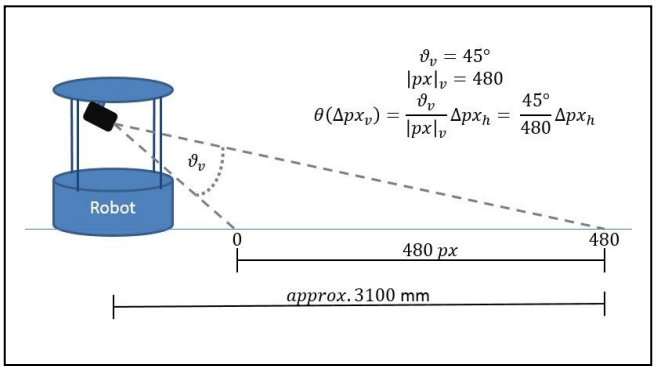


Figure 1. Schematic drawing of the mobile service robot with the Primesense Sensor and the calculation of the vertical aperture angle between two points (Δpx_h).

Therefore, the ground level and the sensor orientation for a better error correction were calculated in the beginning. This step was necessary to prevent the sensor from "losing" orientation between runs or the sensor underlying a drift over time. In this case, a pre-calculated ground plane would lead to a wrong detection of relevant unevenness of the floor.

In a first step, the quality of the current depth image of the sensor is estimated by calculating the RMS deviation of each pixel. For calculating the virtual ground, two points of the middle row and two of the middle column of the depth frame are selected, which satisfy three criteria. The first is that both points have the lowest possible RMS (minimum below the quality factor otherwise use other column or row), the second is a maximized distance between these points and the third criterion is that they do not belong to a known obstacle like walls. This information is taken from the navigation map of the mobile robot platform. In the following section, only the estimation of a vertical ground line is considered because the calculation of the horizontal ground line and also the ground plane is done equally. After the selection of two vertical points, it is possible to calculate the first ground line and the vertical orientation of the sensor. Only five parameters are known: the two distance values of the two selected points, the pixel distance between both points, the vertical aperture angle of the Prime Sense Sensor [52] and the resolution of the current depth frame. Figure 1 shows the aperture angle calculation of each pixel. Together with the pixel distance between the selected points, it is possible to calculate the angle between them. For all examples, a resolution of 640 x 480 pixel is used, which is the highest possible depth resolution of the Prime Sense Carmine Sensor. Using the law of cosines, it is possible to estimate the missing parameters, e.g. the height of the sensor or the vertical angle. After the complete calculation, all relevant values are known in order to be able to estimate the vertical ground line. The next step is similar to the background subtraction. The ground line is a kind of background used for calculating the difference to the current depth image. Figure 2 shows the normal depth image and a binary picture, which is generated by a root-mean-square deviation approach. If the difference is higher than the RMS, the pixel is set to 1, otherwise to 0. Now, it is easier to cluster this picture and find relevant tripping hazards. For clustering,

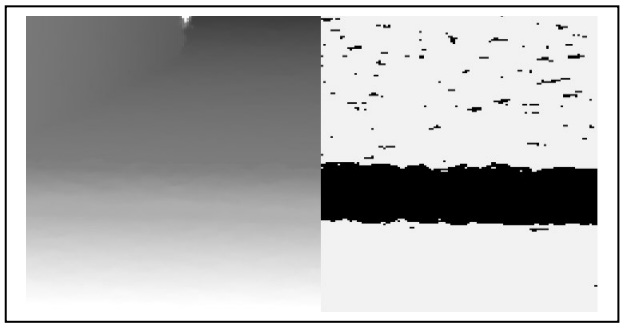


Figure 2. Left side: Depth values from the Sensor in grayscale (White near, dark grey far away) with a 10 mm tread in a distance of 80 cm, right side: Visualization after ground subtraction and converted to a binary image of depth values with the RMS as threshold.

various approaches are published, e.g., edge detection and many more. After found interesting blobs (e.g., size or shape), the height of these obstacles is calculated from the depth picture. This information is saved to the navigation map of the robot. After that, it can be used for scoring the apartment and for increasing precision of gait and balance analysis in the different areas.

Our second approach is similar in respect to the idea that a virtual ground is calculated to use it for a background subtraction and for estimating relevant obstacles on the ground. But instead of calculating the RMS for each pixel, finding the best two pixels near the middle row and column to calculate the virtual ground plane and so on, we used another approach; in respect to the limited calculation power of the mobile robot platform and the gained knowledge of the Prime Sense sensor, only a cut out from the depth image is used. For this approach, the depth picture was taken with the same resolution of 640 x 480 pixel, but only an area of approx. 30 cm vertical and 80 cm horizontal is used, which is located in front of the mobile robot platform. The advantages of this step are:

- four times less pixels in respect to the computational power
- less problems with distant objects, related to a higher sensor noise at greater distances
- higher linearity of the depth image in the area of interest

All these points influence the precision of the ground plane calculation. The disadvantage of a smaller field of view is negligible because of the mobility of the robot platform.

As mentioned in the previous section, only the estimation of a vertical ground line is presented here since the calculation of the horizontal ground line and the ground plane is straight forward. In a first step, a mean depth image of the current location is calculated from over 20 frames, followed by the estimation of median depth values from five columns for each row. The five columns are the middle column and their two left and right neighbors. This new calculated middle median column is used to estimate the horizontal virtual ground line. The first two steps represent a simple filter to reduce the noise of the Prime Sense Sensor due to the limitation of processor capacity. The third step is to calculate a regression line for the middle median column (see (1)). Then the regression line is used as virtual ground line for vertical direction. The following steps are similar to our first approach; the virtual ground is subtracted from the current depth image.

$$RSS = SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)^2 \rightarrow min! \quad (1)$$

If the difference is higher than the median value of the subtraction from the regression line and the middle median column, the pixel is set to 1, otherwise to 0. After finding interesting areas (e.g., size or shape), the height of these areas is calculated from the depth picture. This information is added to the navigation map of the robot. As mentioned in the first approach, these data are used to increase the quality of automated home gait analysis and also for the home score calculation of the HE Assessment.

B. Detection of Unevenness with Triangulation Laser Scanner

As mentioned in Section III.D, it is mandatory to detect an unevenness down to 4 mm, which is near the limit of the most consumer RGB-D Cameras with a detection range from up to 2.5 meters. In order to detect small tripping hazards from 2mm to 20 mm, a triangulation line laser scanner was developed, which consists of two IR line laser modules and a Raspberry Pi B single board computer with associated "NoIR" Camera module, which is able to record visible and infrared light. Both line laser modules are attached to the opposite edges (a distance of 24.5 cm) of the first level of the

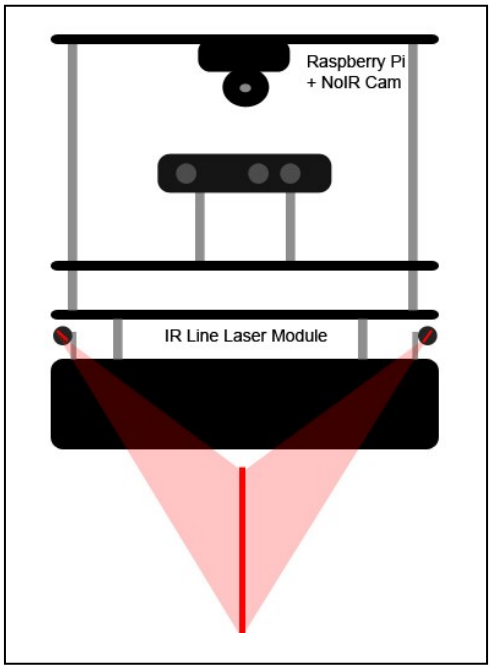


Figure 3. Schematic draw of the set-up of the new triangulation ground laser scanners.

mobile robot platform (at a height of 11.8 cm). The Raspberry Pi with the camera module was mounted upside down to the highest level, approx. 35 cm above the floor (see Figure 3). The camera module can be tilted between 0° to approx. 180°. An angle of approx. 30° was used for the measurements, which provided a horizontal field of view of approx. 36 cm at the beginning and approx. 54 cm at the end and approx. 100 cm in vertical in front of the mobile robot platform. Because of the intensity of the IR laser modules, the entire 100 cm of the vertical field of view were used, which guarantees a good contrast between the IR line and the environment. Both line laser modules were aligned with each other so that they projected a common line onto an even floor in the vertical middle of the camera image, which means that both laser modules have an angle of approx. 46°. The "NoIR" Camera has a single picture resolution of up to 2592 x 1944 pixel [53]. The latest stable OpenCV version 2.4.10 is used for capturing the "NoIR" Camera pictures and the whole computation on each picture.

In a first step, the picture is trimmed to the needed dimensions. As mentioned before, the entire 100 cm of the vertical field of view of the camera were used and have sufficient contrast for the most indoor environments in respect to lighting conditions and floorings. The first step of the image processing is the conversion to a binary image. To find the best possible threshold for this step, two factors are taken into account. The first factor is the current illumination of the room and the second aspect is the current back scatter of the IR line laser on this surface. For the current illumination of a room, the TSL2561 sensor is used.

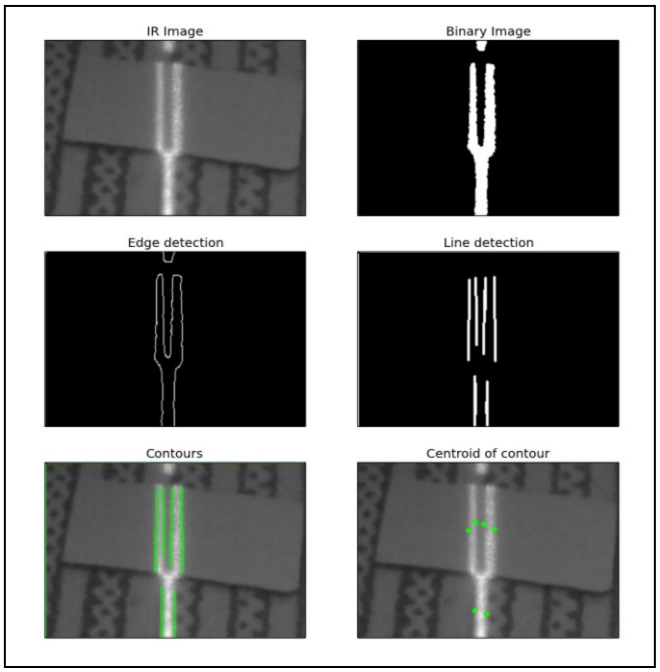


Figure 4. Upper left: part of the raw image from the NoIR Cam with an obstacle (2 mm height), upper right: binary image with calculated threshold, middle left: image after canny edge algorithm, middle right: image after probabilistic Hough line algorithm, lower left: Contours of the Hough lines, lower right: Centroids of the corresponding contours.

A great advantage of this sensor is that it detects both, IR and visible light. So it is possible to estimate the average brightness of the current image without time consuming computation. To improve the estimation, the current reflection of the surface is also taken into account. Therefore, the previous knowledge about the approximate position of the IR laser line within the image is used. In this area, the algorithm searches for the brightest pixel. These two values are used to estimate a threshold value to generate a binary image, in which only the laser line is still visible as a white line. This approach has a high reliability in finding a good threshold to detect only the laser line without having to cut away too much of the edge region of the line. Especially in the upper region of the image, which had - in most cases - the lowest contrast. In the case of the threshold being too high, too much of the edge region has been cut off. This would lead to an inaccurate result during the height estimation of an obstacle. In the case of the threshold being too low, it is possible that there are a lot of artefacts in the binary picture, which make the following computations much more complicated or time consuming.

Afterwards, the OpenCV 2 implementation of the Canny edge detection algorithm with a 5 x 5 kernel is used to find the edges of the laser line. Followed by a probabilistic Hough-Line algorithm, which estimates lines on the base of the canny edge picture (see Figure 4). The result of the probabilistic Hough Line algorithm are different lines, which represent the edges of the canny algorithm. After that, the contours of these lines and their centroids are estimated. Finally, these centroids are used for the calculation of distance and, therefore, for the approximation of the height or depth of obstacles or the unevenness of the floor. To get this final information, some additional steps are needed:

- Finding corresponding centroids
- Calculating distance of corresponding centroids
- Estimate orientation of obstacle
- Calculate level of obstacle

The next step is to find lines that belong to the same segment of a laser line. Therefore, the centroids are sorted depending on the x and y coordinates of each centroid. Together with the pre - knowledge of, e.g., width of a laser line, it is possible to estimate relationships between two centroids. If two pairs of corresponding centroids are found, which are near the same horizontal segment and out of the vertical center, the pixel distance between the both inner centroids of these pairs are calculated. This distance is proportional to the height or depth of an obstacle. To find out if it is a positive or negative elevation of the floor, the left line laser module is switched off and a new image is taken and it is calculated, which lines or centroids are missing. With the knowledge that the even floor is the sectional plane of both line lasers, a left missing line means the obstacle has a positive elevation, a right missing line means the obstacle has a negative elevation. Now there is enough information to calculate the height (positive or negative) and also the length of the obstacle at this point. This information will be added to the corresponding point in the 2D map of the mobile platform. In future developments, it is planned to generate a complete 3D model of this obstacle. Therefore, the fact that

the triangulation line laser scanner is mounted to a mobile robot platform is utilized, and we are able to move the robot along or around an obstacle to estimate the missing parameters, e.g., shape and length. With the additional motion information of the mobile robot platform, it will be possible to generate a complete 3D model of the obstacle.

C. Calculate Balance Parameter

In our first approach, the Microsoft Kinect is used to track the person because of the low price and the existing openNI skeleton tracking algorithm from ROS [54]. The mobile platform does not move during the measurements because of the specification from the openNI algorithm. During the observation phase the timestamp and the x-, y- and z- coordinates of the following skeleton joint point from the openNI tracker node will be saved:

- Foot and hip (each: left, right)
- Torso and Neck

In respect to the low processor capacity of the Turtlebot 2 netbook, an offline approach is used. After the observation phase, different balance and gait analysis parameters are calculated. In a first validation, the distances of the joint points are checked, whether they are between ranges of 0.80 – 3.00m, which is the effective distance of the Kinect sensor. After that, the gait speed, step and stride length and, related to those values, the stance and swing phase of each foot are calculated. First, the different phases for each foot during a measurement are estimated by using (2).

$$\lfloor x_i - x_{i+1} \rfloor_{i=0}^{n} = \begin{cases} \leq 0.02\ m, stance\ phase \\ > 0.02\ m, swing\ phase \end{cases} \quad (2)$$

This means that a foot needs a minimum acceleration of approx. 0.6 m/s to be marked as moving. This value reflects a compromise of literature values and a kind of error correction of the drift from skeleton tracking. After that, the middle index of each phase for each foot is calculated, this is used to estimate stride and step length. Also, the calculation of the gait speed uses these indexes by choosing the first and the last stand phase of each measurement and then calculates the distance between these points. Now, the corresponding timestamps are used to determine the elapsed time. By dividing the distance by time, the gait speed for each measurement is calculated. Two factors are used to get a better reliability between measurements; the first is that the mobile robot stands on a defined OOL, so the global coordinates and the direction are nearly equal between the measurements; the second helpful point is that humans used more or less the same path between two points in the home environment. These points help to get a bigger and comparable data base from the same OOL's

V. RESULTS

In this paragraph the results of our approaches are presented, which were tested and verified in the OFFIS IDEAAL Lab. It provides a complete demo apartment for first measurements in a realistic environment. As a mobile platform, a Turtlebot 2 (Kobuki) was used.

A. Detection of Unevenness with RGB-D Camera

To test and verify our first approach, a Primesense Carmine 1.08 sensor was used, which is mounted upside down underneath the third level of the robot platform and looks down to the ground with an angle of approx. 35 degrees at a height of approx. 34 cm. The resolution of the depth sensor is set to 640 x 480 pixel and a frame rate of 30 Hz. The platform, the sensors and the mounting of both have not been changed during the measurements. To get comprehensive measuring results, the IDEAAL Lab and a normal office space were used to test our approach on different floor types. This configuration gave results from two different carpets, a laminate and a PVC- coating. The measurements in between two floors represent the change between coatings (laminate / carpet). To measure normalized height differences, five wooden tread layers were used. Each piece had a height of 5 mm, so that it was possible to measure between (un-)even doorways (0 mm) up to 25 mm.

For our first approach, we saved 30 single frames for each test set-up, calculated and saved the mean values and the standard deviation for each pixel in order to verify the precision of the sensor. According to different building regulations [48][49][50][51], the requirement is to detect differences of a minimum of 4 mm between two surfaces or an area of 1 square meter. The measured minimal standard deviation is approx. 3.94 mm and the median value is 6.29 mm. This means that the precision of the Prime Sense Carmine 1.08 sensor is near to the required precision of 4 mm. After this result, further tests to verify our first results were performed. Therefore, different measurements in the IDEAAL Lab and within the office with wooden treads were made. The proceeding for each measurement was the same; first, 30 frames of the even surface were taken, then 30 frames with a 5 mm tread in a distance of 80 cm followed by 30 frames with 10 mm tread and so on until the maximum of 25 mm was reached. After that, the distance was reduced to 40 cm and started over without any obstacles and then raised the treads in 5 mm steps.

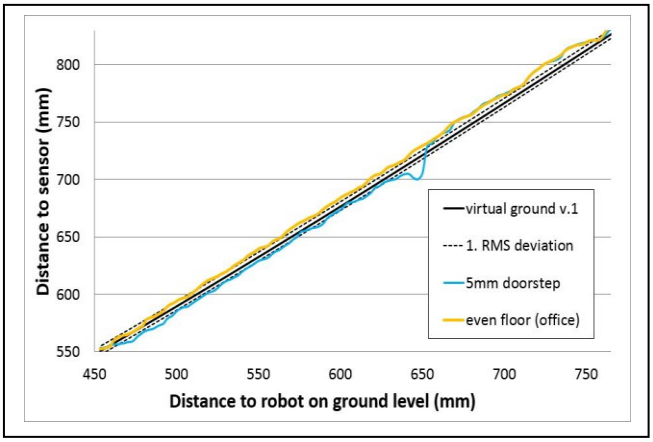


Figure 5. Visualization of the calculated virtual ground v.1 (black), the first RMS deviation (dotted lines) and the measurement from the ground (orange) and a 5 mm tread (blue).

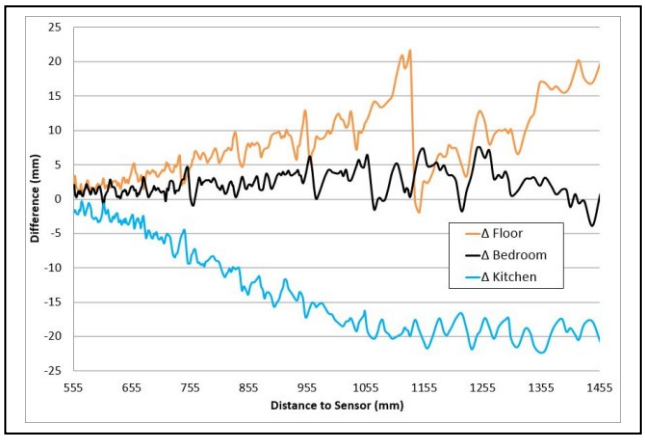


Figure 6. Visualization of dependency of different floorings in comparison to the general mean ground value.

After the measurement, the virtual ground plan was calculated and subtracted from different test images. The result was unexpected; in the first approach, only two small areas had good results. These areas were around the selected points for the calculation of the ground plane. Even for a floor without any unevenness. After a small modification (also considered in the description of the approach) of the algorithm, which selects the point for estimate the ground plane, a vertical ground line was calculated, which only matched the lower third of the depth picture. Figure 5 shows that the difference between the calculated ground and the real ground in the upper two thirds of the picture was too big to detect any relevant barriers.

After these results, the first step was to verify the measurement, by subtracting the mean value of the even ground from the mean values of the modified ground. These values showed acceptable results for the detection of barriers from 5 mm up to 25 mm. The next step was the linearity of the sensor over distance. If it had a linear characteristic for the depth sensor, then our approach should work in general. The result in Figure 6 shows that the sensor does not have a perfect linear characteristic with objects more than 1.4 m away from the sensor. This means that our approach to calculate a virtual ground, which is represented by a plane or line and use it for a simple background subtraction, is not applicable to the complete range of the Prime Sense sensor. After that, our second approach was developed and tested with the same data set, which we generated during the test for our first approach. This guaranteed a high comparability of these two approaches. As mentioned in Section IV.A, only a cut out of the original depth images was used for the second approach. The final depth image for the second approach has a size of 320 x 240 pixel, which represented the lower half of the original picture, it belongs to an area of approx. 30 cm (vertical) x 80 cm (horizontal) in front of the mobile robot platform. Based on this extracted data set, the new virtual ground was calculated with (1). The virtual ground was subtracted from depth images of all 12 set-ups. The results for our second approach were much better than for the first approach. As you can see in Figure 7, the new virtual ground fits nearly perfectly to the depth image of the even floor. Also, the difference to the depth image with the 5 mm tread seems to be good enough to be able to guarantee a detection of obstacles of at least 5 mm. Now, the second approach was tested if the main goal of being able to recognize obstacles in an easy way without having a complete 3D map of the even floor in an apartment could be achieved. Therefore, the measurement was repeated in different rooms of the IDEAAL apartment with different floor types and subtracted the new depth images from the generated virtual ground v.2. These results were surprising again. In some rooms, the virtual ground v.2 fits quite well to the depth image of the even floor and the differences are in the first order of the RMS deviation. But in some cases, huge areas were found, which were marked as potential obstacles on an even floor. A good example was the depth image from the kitchen (see Figure 8). In the lower area, both lines fit quite well but the in the upper area the depth image and the virtual ground v.2 have a great gap. The difference between the calculated and the real ground is bigger than the first RMS deviation, which means that false positive barriers were detected. The difference between the virtual ground v.2 and a 5 mm tread in the kitchen is only few mm above the first standard deviation. Also the difference between the even ground and the virtual ground v.2 are quite better than the difference between the even ground and virtual ground v.1 (see Figure 9). It seems to be possible to detect obstacles with a height of approx. 10 mm but as mentioned in Section III.D for the HE Assessment, obstacles needed to be detected with a height of 4 mm. Therefore, a sensor resolution up to 1.5 – 2 mm is needed.

B. *Detection of Unevenness with Triangulation Laser Scanner*

Because of the insufficient results from our two approaches with the RGB-D-Camera, an own triangulation laser scanner was developed as described in Section IV.B. In order to evaluate this new scanner, it was tested under comparable conditions to the approaches with the Prime Sense Carmine sensor. In a first test, the office floor with different obstacles was used. These had a height of 2 mm, 5 mm, 10 mm, 15 mm, 20 mm and 25 mm and were placed in front of the mobile robot platform in a distance of approx. 80 cm and 40 cm, which is similar to the set-up of our first two approaches. For each set-up, 10 pictures were taken with the Raspberry Pi NoIR Camera for estimating the height of each obstacle with the approach of Section IV.B. As you can see in Figure 4, both laser lines are clearly separated, also for obstacles of a height of down to 2 mm. This depends on a relative low position over ground and great distance between both laser modules, which resulted in a shallow angle and guaranteed a good separation for low obstacles. Because of the relatively high position of the Raspberry Pi and its camera module, we have a great field of view and so it is also possible to detect obstacles up to 25 mm. The upper right picture in Figure 4 shows only a cut out from the whole image of the NoIR Camera module. The obstacle that is shown was placed in front of the mobile robot platform with a distance of approx. 80 cm and the picture still shows a good resolution and contrast. Therefore, it is possible to

estimate the height with an accuracy of approx. 1.5 mm over the complete range of 100 cm.

C. Dependence of the Surface

The dependence of the Prime Sense sensor on the flooring was tested by measuring four different floor types, two different kinds of carpet, PVC- coating and laminate. Also, the transition from laminate to carpet was tested. For each surface, 30 single measurements were made and the mean value over all 30 single frames on pixel base was computed. Then, these mean values were used to calculate the overall mean value of the ground. The mean value of the middle column was selected from each measurement and subtracted from the corresponding value of the overall mean depth picture. The results are shown in Figure 6 and lead us to the fact that different floorings have an influence on the distance values and the reliability of the sensor. As you can see in Figure 6, the deviations in the first 50 pixel, which are equivalent to a distance of approx. 20 cm in front of the mobile robot platform, represent a difference within the first RMS deviation of about 3.94 mm. The total measurement represents a distance between approx. 10 cm to 84 cm from the mobile robot platform. This result points out that it is advisable to calibrate the sensor daily and for each subsurface in order to reduce errors during the measurement, or use a different model of this sensor type, e.g., the Prime Sense Carmine 1.09 with higher depth resolution or a complete other type of sensor to detect the unevenness of the floor. We also conducted first tests on different floorings with our triangulation laser scanner. Right now only two different set-ups have been tested; on the office carpet and the PVC in the IDEAAL apartment kitchen. The line laser modules obtain an energy output of approx. 5mW, which should guarantee a good contrast over a wild field of different variables (i.e., sunlight, floor type). The first test was to estimate a threshold for the office carpet, which was relatively easy because of the good optical properties. This means, the back scatter of the IR light was very high and the picture had a good contrast. So, it was not a challenging test for our threshold estimation because the range for a good

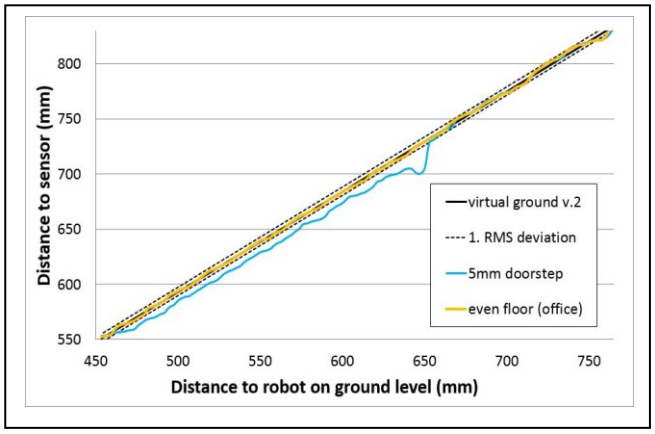


Figure 7. Visualization of the calculated virtual ground v.2 (black), the first RMS deviation (dotted lines) and the measurement from the ground (orange) and the 5 mm tread (blue).

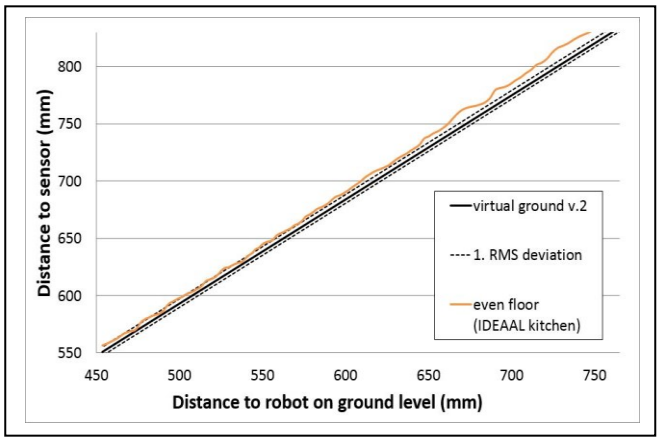


Figure 8. Visualization of the calculated virtual ground v.2 (black), the first RMS deviation (dotted lines) and the measured ground of the kitchen (orange).

threshold was relatively great. As a second test flooring, a PVC-coating was used, which laid in the IDEAAL apartment kitchen. It has a shiny finish, which means that the back scatter is lower, which makes it challenging to our approach of finding the right threshold. Under some circumstances, e.g., sunny days, it was possible that the estimated threshold was too low in order to be able to separate the laser line from the rest. The result was that the binary picture had some sprinkles. Most of these artefacts were too small / short, so that the canny edge algorithm or the probabilistic Hough line algorithm ignored them. But in few cases they lead to false positive centroids. During the correspondence check, it was possible to separate these points. For final results, more tests have to be performed to be sure that our new triangulation line laser scanner works reliably under most conditions in an apartment and that it is not possible that sprinkles lead to wrong results under special circumstances.

D. Gait Parameters vs. Floorings

Parallel to the tests for detection of unevenness of the ground, first measurements in a domestic environment with five users (two females/ three males) between 42 – 76 years were made. These results are used as a first validation of our approach for calculating gait speed, stride and step length and, when possible, to see differences between different floorings by using the Microsoft Kinect and openNI tracker. For all measurements, the Turtlebot 2 stands at a predefined position, similar to the final setup when the mobile robot drives to various OOL's for measurement. Each subject had to walk towards the mobile robot five times under the same conditions. Each test person had to fulfill this test with 10 different conditions. In general, they had to walk over two different coatings (carpet / parquet). On each coating, three treads of different height (5 mm, 10 mm and 25 mm height) were placed in the middle of the walking distance. The test person also had to manage all these set-ups under dark and normal lighted conditions. This lead to a data base of 250 measurements including all conditions and subjects. The first results for the step-, stride length and self-selected gait speed (SGS) on parquet, high pile carpet and different treads are

presented. As depicted in Figure 10 and Figure 11, a difference between stride length and SGS could not be detected for elderly persons only but also for mid-aged persons, depending on the floorings. Also, it seems as if the variation of step- and stride length depends on the coatings. But further tests with more measurements, longer walking distances and time periods must be conducted to verify our first results. Nevertheless, evidence that the floorings have an impact on the gait analysis in the domestic environment was shown. Without the knowledge of the characteristics of the flooring, like the most classical automated approaches, it could lead to false decisions related to the decreasing of the SGS on some coatings. This gives first evidence that the quality of balance and gait analysis depends also on the floorings. Further tests must be conducted in order to get reliable data on what kind of obstacles have an influence and how big the impact is.

VI. Conclusion and Future Works

A new approach for the detection of fall relevant unevenness of floorings and a first idea of an advanced gait analysis, which uses this information for enhanced results in the context of an automated HE Assessment was presented. For this, a mobile robot platform, i.e., the Turtelbot 2 was used. As a depth sensor, a Prime Sense Carmine 1.08 with the original OpenNI driver v.2.1.0 and a self-constructed triangulation line laser scanner was used for the detection of unevenness; and a Microsoft Kinect with the ROS openNI tracker Node was used for the balance and gait analysis. The Carmine sensor was mounted up-side down underneath the third level of the Turtelbot platform in a height of approx. 34 cm. For the triangulation line laser scanner approach, the Prime Sense sensor was replaced by the Raspberry Pi with the NoIR camera module. The Kinect was mounted to the highest level (height approx. 55 cm). Our approach with the RGB-D camera aimed at a calculated virtual ground, which is the reference for barriers because in a normal scenario it is unrealistic to have the chance to make a clean depth picture from each part of the room without any carpets on the subsurface or other stumbling blocks. It was possible to determine the position and orientation of the sensor only from a small knowledgebase. Still, our measurements have shown that the combination of our approach with this sensor, the mounting and the needed resolution does not work in a proper way.

This depends on three factors:

- First: the depth resolution of the sensor. The noise of the sensor is near the values that we want to detect.
- Second: the dependence of the sensor. As shown, the floorings and the gloss of it have a big influence on the depth values. The difference is sometimes even more than the third standard deviation.
- Third: the quality of our first algorithm to select the points for the calculation of the virtual ground.

Our second approach shows better results for the estimation of the virtual ground (see Figure 9) but the first

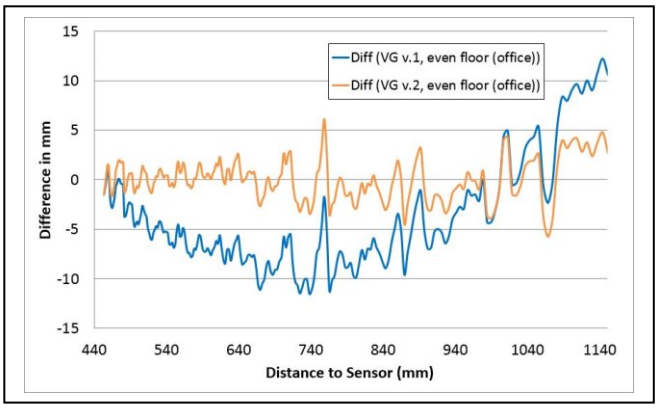


Figure 9. Shows a comparison between our first (VG v.1, blue) and second (VG v.2, orange) approach of estimate a virtual ground. The difference between the calculated virtual grounds and the measured ground of the office is shown.

two points are still valid. There are different possibilities to cope with these problems; it is possible to try better filter algorithms over more frames to reduce the noise and get better results or try to generate different virtual grounds for each room to handle the dependence on different floorings. But this step would lead us away from our original idea of having a general virtual ground. Finally, we could say that the Prime Sense Carmine 1.08 sensor has some advantages, like the price, the relatively good resolution and low noise in relation to the price and range. But the quality is not high enough for this application in the frame of HE Assessment or to determine relevant unevenness of the ground. Our triangulation line laser scanner shows better results in respect to accuracy for the estimation of the height of obstacles.

The dependency on different types of surfaces seems to be lower compared to the Prime Sense Carmine 1.08 sensor. For general valid answers, we have to conduct more tests with this new sensor. In general, this sensor type has the disadvantage of generating 2D information only, since it can only analyze a height profile along a single line. In a further step, the mobile robot platform will be used to generate 3D information by moving the sensor to different points but the computational demands are still higher compared to a RGB-D camera, which generates an entire 3D point cloud of the surrounding.

Our approach to use additional information on the floorings in order to raise the quality of gait analysis in the domestic environments seems to be essential to generate reliable data. As a first result, we were able to show that an influence of the flooring exists but for final statements we have to evaluate this approach with more users and with more flooring and other influence factors. The first results allow the statement that all automated gait analyses in unsupervised environments should consider the texture and unevenness of the flooring.

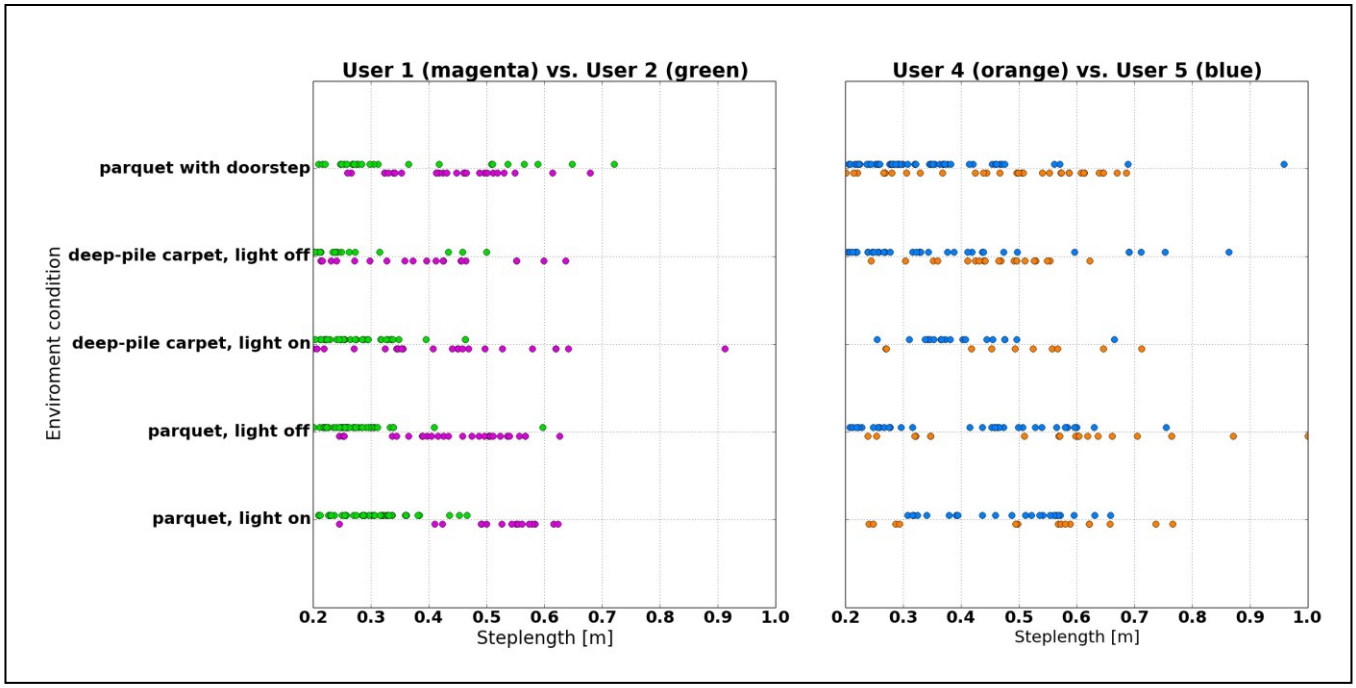


Figure 10. Influence of floor conditions to the step-length of different subjects (magenta/orange: mid-age, green/blue: elderly). Left side: two female subjects and on the right site two male subjects.

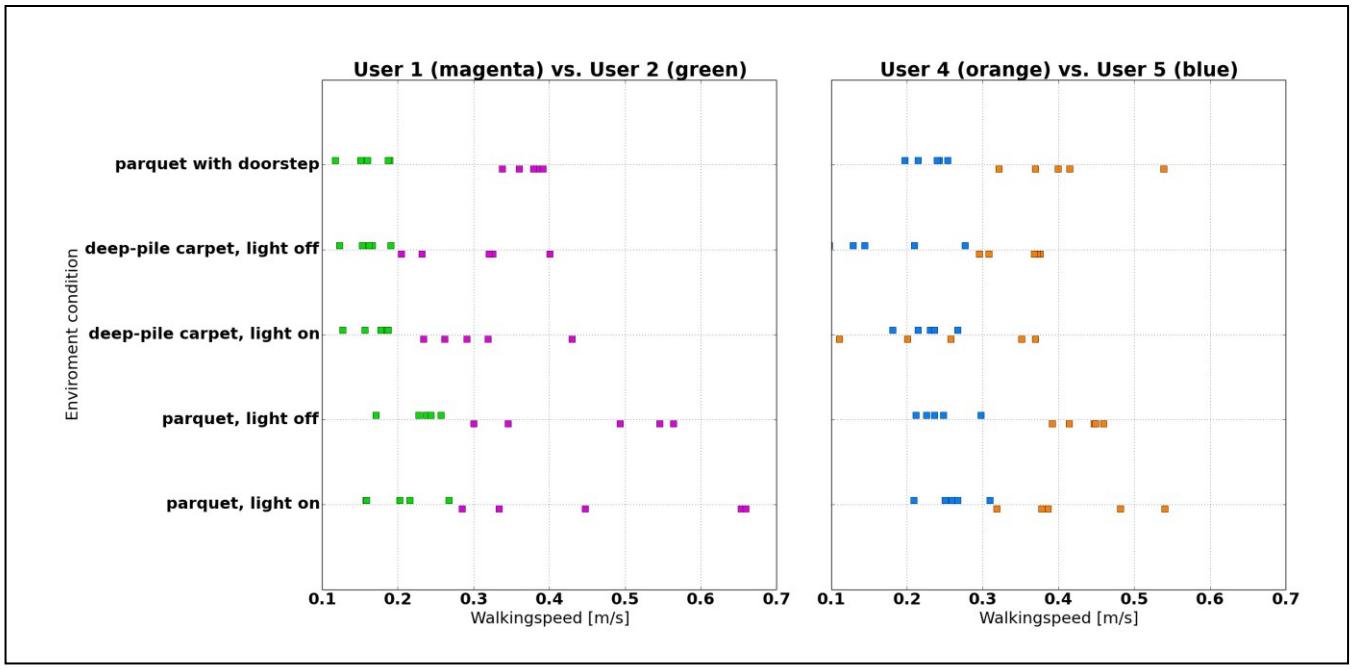


Figure 11. Influence of floor conditions to the gait speed of different subjects (magenta/orange: mid-age, green/blue: elderly). Left side: two female subjects; Right site: two male subjects.

REFERENCES

[1] N. Volkening and A. Hein, “Using an Autonomous Service Robot for Detection of Floor Level Obstacles and its Influence to the Gait,” The Fourth International Conference on Ambient Computing, Applications, Services and Technologies (Ambient 2014) IARIA, Rome, August 24, 2014, pp. 65-71, ISBN: 978-1-61208-356-8.

[2] K. Böhle, K. Bopp, and M. Dietrich, “The "Artificial Companion" - a useful guiding principle for development and implementation of technical assistance systems in care arrangements?,” In Proceedings of: 6. German AAL-Congress: " Quality of life in change of demographics and technology " VDE, Berlin, 2013, ISBN: 978-3-8007-3484-9.

[3] D. J. Cook and S. K. Das, “How smart are our environments? An updated look at the state of the art,” Pervasive and Mobile Computing, vol. 3, no. 2, 2007, pp. 53 – 73.

[4] J. M. Beer, A. Prakash, T. L. Mitzner, and W. A. Rogers, “Understanding Robot Acceptance,” Technical Report HFA-TR-1103 Atlanta, 2011, GA: Georgia Institute of Technology School of Psychology – Human Factors and Aging Laboratory, Online available: https://smartech.gatech.edu/handle/1853/39672, last access: 2015.05.26.

[5] F. Vaussard et al., “Lessons Learned from Robotic Vacuum Cleaners Entering in the Home Ecosystem,” Robotics and Autonomous Systems, Volume 62, Issue 3, March 2014, pp 376–391.

[6] J. Forlizzi, “How Robotic Products Become Social Products: An Ethnographic Study of Cleaning in the House,” Human-Robot Interaction (HRI), 2007 2nd ACM/IEEE International Conference on, Arlington, VA, 9-11. March 2007, pp. 129-136, ISBN: 978-1-59593-617-2.

[7] J. Meyer, M. Brell, A. Hein, and S. Gessler; “Personal Assistive Robots for AAL Services at Home - The Florence Point of View,” 3rd. IoPTS workshop, Brussels, 2009.

[8] T. Rehrl et al., “The Ambient Adaptable Living Assistant is Meeting its Users,” AAL Forum 2012, 24 - 27 September, Eindhoven, Netherlands.

[9] T. Frenken, M. Isken, N. Volkening, M. Brell, and A, Hein, “Criteria for Quality and Safety while Performing Unobtrusive Domestic Mobility Assessments Using Mobile Service Robots,” Ambient Assisted Living, Advanced Technologies and Societal Change 2012, 5. AAL-Congress 2012 Berlin (VDE), Germany, 24-25 January, 2012, pp. 61-76, doi: 10.1007/978-3-642-27491-6_5.

[10] N. Volkening, A. Hein, M. Isken, T. Frenken and M. Brell, “HE – Detection of imminent risk areas in domestic environments using mobile service robots,” 6. German Ambient Assisted Living Congress, Berlin, Germany, VDE, 2013, pp. 479-485.

[11] D. G. Bruce, A. Devine, and R. L Prince, “Recreational Physical Activity Levels in Healthy Older Women: The Importance of Fear of Falling,” Journal of the American Geriatrics Society, Volume 50, Issue 1, pages 84–89, January 2002, doi: 10.1046/j.1532-5415.2002.50012.x.

[12] K. Balzer et al., „Prevention of falls for older people in their own home environment,“ Health Technology Assessment 116, 2012, online: http://portal.dimdi.de/de/hta/hta_berichte/hta255_bericht_de.pdf, last access: 2015.05.26.

[13] W. von Renteln-Kruse et al., “Medicine of aging and older people,” journal of Gerontology and Geriatrics, Volume 38, Issue 4, pp 288-292, August 2005, doi: 10.1007/s00391-005-0274-1.

[14] A, Kochera, ”Falls Among Older Persons and the Role of the Home: An Analysis of Cost, Incidence, and Potential Savings from Home Modification,” AARP Public Policy Institute, March 2002, online: http://assets.aarp.org/rgcenter/il/ib56_falls.pdf, last access: 2015.05.26.

[15] F. J. Imms and O. G. Edholm, “Studies of gait and mobility in the elderly,” Age Ageing, vol. 10, no. 3, August 1981, pp. 147–156.

[16] M. Montero-Odasso et al., “Gait Velocity as a Single Predictor of Adverse Events in Healthy Seniors Aged 75 Years and Older,” Journal of Gerontology: Medical Sciences, vol. 60, no. 10, October 2005, pp. 1304–1309, doi: 10.1093/gerona/60.10.1304.

[17] R. Cham and M. S. Redfern, “Changes in gait when anticipating slippery floors,” Gait and Posture 15, pp. 159-171, 2002.

[18] S. B. Thies, J. K. Richardson, and J. A. Ashton-Miller, “Effects of surface irregularity and lighting on step variability during gait: A study in healthy young and older women,” Gait and Posture, vol. 22, 1. August 2005, pp. 26-31, ISSN 0966-6362.

[19] D. S. Marigold and A. E. Patla, “Age-related changes in gait for multi-surface terrain,” Gait and Posture, vol. 27, no. 4, May 2008, pp. 689-696.

[20] C. N. Scanaill et al., “A Review of Approaches to Mobility Telemonitoring of the Elderly in Their Living Environment,” Annals of Biomedical Engineering, vol. 34, no. 4, April 2006, pp. 547–563.

[21] K. Cameron, K. Hughes, and K. Doughty, “Reducing fall incidence in community elders by telecare using predictive systems,” in Proc. 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 3, 1997, pp. 1036–1039, ISSN :1094-687X.

[22] J. A. Kaye et al., “Intelligent Systems for Assessing Aging Changes: Home-Based, Unobtrusive, and Continuous Assessment of Aging,” The journals of gerontology. Series B, Psychological sciences and social sciences, vol. 66, iss. 1, 1. July 2011, pp. i180–i190, doi: 10.1093/geronb/gbq095.

[23] M. P. Poland, D. Gueldenring, C. D. Nugent, H. Wang, and L. Chen, “Spatiotemporal Data Acquisition Modalities for Smart Home Inhabitant Movement Behavioural Analysis,” ICOST '09, Proceedings of the 7th International Conference on Smart Homes and Health Telematics, Springer, 2009, pp. 294-298.

[24] T. Frenken et al., “A novel ICT approach to the assessment of mobility in diverse health care environment,” CEWIT-TZI-acatech Workshop, "ICT meets Medicine and Health" (ICTMH 2013), April 2013.

[25] T. Pallejà, M. Teixidó, M. Tresanchez, and J. Palacin, “Measuring Gait Using a Ground Laser Range Sensor,” Sensors, vol. 9, no. 11, 2009, pp. 9133–9146.

[26] E. E. Stone and M. Skubic, “Passive In-Home Measurement of Stride-to-Stride Gait Variability Comparing Vision and Kinect Sensing,” 33rd Annual International Conference of the IEEE EMBS, Boston, Massachusetts, USA, 2011, pp. 6491-6494.

[27] M. Gabel, R. Gilad-Bachrach, E. Renshaw, and A. Schuster, “Full Body Gait Analysis with Kinect,” 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), San Diego, USA, 2012, pp. 1964-1967.

[28] E. E. Stone and M. Skubic, “Unobtrusive, Continuous, In-Home Gait Measurement Using the Microsoft Kinect,” IEEE Transactions on biomedical engineering, vol. 60, no. 10, October 2013, pp. 2925-2932.

[29] A. Petrovskaya and A. Y. Ng, “Probabilistic mobile manipulation in dynamic environments, with application to opening doors,” in International Joint Conference on Artificial Intelligence (IJCAI), 2007, pp. 2178-2184.

[30] C. L. Breazeal, “Sociable machines: Expressive social exchange between humans and robots,” Ph.D. dissertation,

Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2000.

[31] C. Ray, F. Mondada, and R. Siegwart, "What do people expect from robots?," in IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008, pp. 3816–3821.

[32] P. Udsatid, N. Niparnan, and A. Sudsang, "Human Position Tracking for Side By Side Walking Mobile Robot using Foot Positions," Proceedings of the 2012 IEEE International Conference on Robotics and Biomimetics, 11-14. December 2012, pp. 1374 – 1378, Guangzhou, China.

[33] M. Brell, J. Meyer, T. Frenken, and A. Hein, "A Mobile Robot for Self-selected Gait Velocity Assessments in Assistive Environments," in The 3rd International Conference on Pervasive Technologies Related to Assistive Environments (PETRA'10), Samos, Greece, June 2010, ISBN 978-1-4503-0071-1.

[34] M. Isken et al., "Enhancing Mobile Robots' Navigation through Mobility Assessments in Domestic Environments," in Proceedings 4. German Congress, Ambient Assisted Living, VDE Verlag, 2011, pp. 223-238.

[35] T. M. Gill et al., „A Population-Based Study of Environmental Hazards in the Homes of Older Persons," American Journal of Public Health, Vol. 89, No. 4, April 1999, pp. 553-556,

[36] M. E. Northridge, M. C. Nevitt, J. L. Kelsey, and B. Link, "Home Hazards and Falls in the Elderly: The Role of Health and Functional Status," American Journal of Public Health, Vol. 89, No. 4, April 1999, pp 509-515.

[37] G. Carlsson, B. Slaug, A. Johannisson, A. Fänge, and S. Iwarsson,. "The Housing Enabler - Integration of a computerised tool in occupational therapy undergraduate teaching," CAL Laborate, June, 2004, pp. 5 – 9.

[38] A. Fänge, "Strategies for evaluation of housing adaptations – Accessibility, usability and ADL dependence," ISBN: 91-974281-5-9. doctoral thesis, Department of Clinical Neuroscience, Lunds University, Lund, Sweden, 2004.

[39] T. Helle et al., "The Nordic Housing Enabler: Inter-rater reliability in cross-Nordic occupational therapy practice," Scandinavian Journal of Occupational Therapy, 17. December 2010, pp. 258-266.

[40] M. Cesari et al, "Prognostic Value of Usual Gait Speed in Well-Functioning Older People—Results from the Health, Aging and Body Composition Study," Journal of the American Geriatrics Society, vol. 53, 2005, pp. 1675–1680.

[41] S. E. Carter, E. M. Campbell, R. W. Sanson-Fisher, S. Redman, and W. J. Gillespie, "Environmental hazards in the homes of older people," Age and Ageing, 1997; vol. 26, pp. 195-202.

[42] R. Cham, and M. S. Redfern, "Changes in gait when anticipating slippery floors," Gait and Posture, Vol. 15, 2002, pp. 159– 171.

[43] S. B. Thies, J. K. Richardson, T. Demott, and J. A. Ashton-Miller, "Influence of an irregular surface and low light on the step variability of patients with peripheral neuropathy during level gait," Gait Posture, vol. 22, August 2005.

[44] Y.-J. Yu, I.-S. Shin, K.-K. Lee, T.-J. Yoon, C.-S. Choi, and C.-S. Chung, "A kinematic analysis of elderly gait while stepping over obstacles of varying height," XXV ISBS Symposium 2007, Ouro Preto – Brazil, ISSN 1999-4168, online available: https://ojs.ub.uni-konstanz.de/cpa/article/view/565/504, last access: 2015.05.26.

[45] H. C. Chen, J. Ashton-Miller, N. B. Alexander, and A. B. Schultz, „Stepping over obstacles: gait patterns of healthy young and old adults," Journal of Gerontology, November 1991, 46(6):M196-203.

[46] B. J. McFadyen and F. Prince, "Avoidance and accommodation of surface height changes by healthy, community-dwelling, young, and elderly men," Journal of Gerontology: Biological Sciences, April 2002, 57(4):B166-174.

[47] J. Sun, M. Walters, N. Svensson, and D. Lloyd, „The influence of surface slope on human gait characteristics: a study of urban pedestrians walking on an inclined surface," 1996, Ergonomics, 39:4, pp. 677-692, DOI: 10.1080/00140139608964489.

[48] Professional association rules for safety and health at work, BGR 110, April 2007, Federation of Trade Associations, Online: http://publikationen.dguv.de/dguv/pdf/10002/bgr-110.pdf, last access: 2015.05.26.

[49] Department of Environmental Health and Safety, Stanford University, "Slip, Trip, and Fall Prevention Guide," January 2008, online available: https://web.stanford.edu/dept/EHS/prod/mainrencon/occhealth/slip_trip_fall_prevention.pdf, last access: 2015.05.26.

[50] Joint ministerial order, "Technical Regulations for Workplaces ASR A1.5/1,2," GMB, February 2013, pp 348ff, online available: http://www.baua.de/de/Themen-von-A-Z/Arbeitsstaetten/ASR/pdf/ASR-A1-5-1-2.pdf, last access: 2015.05.26.

[51] DIN 18202:2013-04, "Tolerances in building construction – Buildings," German Institute for Standardization e.V, 2013.

[52] Primesense – 3D Carmine 1.09 Sensor, Product Information, Available Online: http://i3du.gr/pdf/primesense.pdf, last access: 2015.05.26.

[53] Raspberry PI NoIR Camera Documentation, Online available: http://docs-europe.electrocomponents.com/webdocs/127d/0900766b8127db33.pdf, last access: 2015.05.26.

[54] Zhengyou Zhang, "Microsoft Kinect Sensor and Its Effect," IEEE Multimedia, vol. 19, no. 2, pp. 4-10, February 2012, doi:10.1109/MMUL.2012.24.

Decision Support System for Neural Network R&D

Rok Tavčar
and Jože Dedič
Cosylab, d.d.
Ljubljana, Slovenia
Email: rok.tavcar@cosylab.com,
joze.dedic@cosylab.com

Andrej Žemva
Faculty of Electrical Engineering
University of Ljubljana
Ljubljana, Slovenia.
Email: a.zemva@ieee.org

Drago Bokal
Faculty of Natural Sciences and Mathematics
University of Maribor
Maribor, Slovenia
Email: drago.bokal@uni-mb.si

Abstract— **One of the reasons that keep Neural Networks (NNs), which are advanced computational methods (ACM) of great potential, from coming into broader practical use, is the lack of systematic method in finding the optimal match between NN architecture and target application. If this match is performed erratically, practical solutions often yield unimpressive results. It is the a) validation of the problem's fitness for a NN-based solution and b) matching of an optimal NN implementation to the given problem that is crucial. This paper presents a theoretical foundation for an inference engine decision space and a taxonomic framework for a knowledge base, which are part of our proposed knowledge-driven decision support system (DSS). Furthermore, this paper provides details of our inference engine, namely a) an algorithm for optimal matching of a NN setup against the given learning task, b) the application of this same algorithm for interactive exploration of the decision space and c) an algorithm for automatic inference of potential NN research synergies based on existing successful NN solutions. Finally, we propose a process for establishing and maintaining the growth of the DSS knowledge database.**

Keywords—Neural Networks; DSS; Knowledge Base; Taxonomy; Inference Engine.

I. Introduction

This paper presents a theoretical foundation for an inference engine decision space and a taxonomic framework for a knowledge base, which are part of our proposed knowledge-driven DSS. Furthermore, it introduces new additions to the suite of related algorithms, previously presented in [1].

The compelling notion, that NNs are universal approximators [2], leads quickly to believe, that any NN will do well on any presented machine learning task. However, the universal approximation theorem only guarantees the existence of an approximation, but not that it can be learned, nor that it would be efficient. Practice shows that every given problem requires a carefully crafted NN design and that advanced NN concepts, tailored to specific types of tasks are necessary to attain best results. This factor, among others, has led to the existence of a large number of conceptually varying NN architectures and learning algorithms [3].

For best results, any researcher or practitioner of today needs to understand a vast domain of knowledge in order to find a NN solution most suitable to their task. Due to domain vastness, researchers often limit themselves to NN domains they are familiar with, preventing new knowledge from propagating efficiently among all who would benefit from it. Specifically, we thus face a twofold handicap for progress of NN research: a) practitioners use suboptimal NN setups for real-world applications [3], inhibiting broader NN acceptance in the industry and b) researchers delve into local extrema of research (e.g., through jumping on the bandwagon of imminent peers [4]), pushing frontiers of NN research in suboptimal directions.

Figure 1 shows a simple flowchart view of the current typical approach to selection of NNs for chosen learning task. It illustrates that the lack of systematic approach to NN selection often yields suboptimal results. This has a negative effect on a wider acceptance of NNs in the industry. A key prerequisite in current NN design is expert intuition, which can be attained either through significant experience with NN implementation and applications in practice, or through access to expert intuition in an environment of experienced NN users. When expert intuition is present in early stages of design, the subsequent efforts give promising results (Figure 1, left); and the opposite, when not (Figure 1, right). The NN community needs a streamlined way of enabling existing and potential NN users to make optimal technological choices efficiently and systematically. Having today's foremost NN research applied in the industry can foster wider acceptance of NNs into practice and improve NN research.

In 2006, Taylor and Smith [5] created an important taxonomy-based evaluation of NNs, which aids in validating whether a given problem is solvable with a NN at all. The next concern, which they point out and we hereby address, is to choose the right NN architecture and its concrete implementation for the problem. Our goal is to provide a DSS for industry practitioners and researchers to systematically find the right NN for their application or research interest. This paper proposes a solution that enables (1) a systematical overview of the complete NN knowledge domain to (2) compare NN instances through their capabilities in a (3) quickly interpretative way using a framework that is (4) adaptable in terms of NN properties, even classification dimensions.

This paper is structured as follows. Section II briefly outlines the state of the art, Section III explains our approach to DSS design. Sections IV and V present the DSS' inference engine decision space and the taxonomy for DSS' knowledge

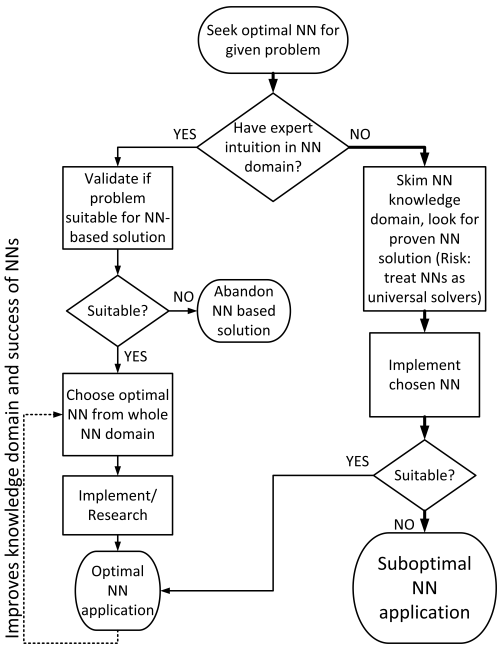


Figure 1. Typical design flow of selection and implementation of NNs for a given learning task.

base, respectively. Sections VI and VII present the DSS and its main use cases, respectively. Section VIII proposes strategies for maintenance of the knowledge database, and finally, Section IX concludes this paper and outlines the future work.

II. State of the Art

The state of the art in NN design methodology can be split into two groups. The first group focuses on choosing the optimal NN *macrostructure* (e.g., Support Vector Machine versus Recurrent Neural Network), while the second group focuses on guidelines and (semi)automated methods, that help find the optimal *microstructure* (e.g., number of hidden layers and neurons) of a selected macrostructure.

The first group of approaches consists of guidelines and overview literature [6][7][8]. The problem (and virtue) with this set of methodologies is that they require understanding of a vast set of NN concepts, before the designer is able to make an optimal choice. Dreyfus [6] states, for example: "No recipes will be provided here. It is our firm belief that no significant application can be developed without a basic understanding of the principles and methodology of model design and training." Of course, we agree with this position. However, it can be observed in practice, that there is a lack of systematic approach in choosing the macrostructure. As a consequence, Feedforward NN (FFNN) [2], learned with Backpropagation (BP) [9], is still chosen in the majority of applications, which we consider a negative trend [10].

The second group of approaches is necessary for the fine microstructural tuning of a chosen macrostructure (also usually demonstrated on FFNN with BP). These approaches are either given as a set of rules and recipes, or as an automated optimization tool. The most systematic approaches rely on the Design of Experiments (DoE) method, involving Taguchi principles [11][12][13]. Such methods systemize and automate the selection of, e.g., number of hidden layers or neurons, through experimenting with different setups. Similar methods are constructive and pruning algorithms, that add or remove neurons from an initial architecture [14][15]. Also, related are evolutionary strategies, which employ genetic operators for similar purposes [16][17][18].

We turn to the state of the art in search of a formal taxonomy of the whole NN knowledge domain. For a taxonomy to serve as the foundation of our DSS, it must facilitate a qualitative measure between its elements. However, it turns out, that directly comparing NN instances in detail is prohibitively problematic due to bias or lack of method in the description process [19]. The Andrews-Diederich-Tickle (ADT) taxonomy [20] enables two NNs to be compared pairwise through ADT^5 criteria (defined by Andrews et al. [20] and refined by Tickle et al. [21]), but this taxonomy lacks orthogonality since some of its taxonomical categories (dimensions) are interdependent. Other taxonomies classify NNs purely through topology [22] or realization method [9]. And more recently, researchers create taxonomies that assist in choosing the best solution for the task [4][23] within a limited application area and solve locally what our work attempts to solve globally.

III. Design of the Decision Support System

Our DSS-based approach proposed fits between the two groups of NN design methodology, presented in Section II, and improves the results of both group's goals. It exhibits the main qualities of the second group (ability to automate the decision process) and applies them to the problematic of the first group (i.e., choosing the macrostructure), which is a crucial step in NN design, because the effect of any design actions depends greatly on early decisions. The aim of our proposed DSS is to improve the performance of NN-based based applications on a large scale, through enabling designers to perform optimal early design decisions. Figure 2 illustrates how our proposed DSS improves the NN design process by enabling users to systematically find optimal NN instances for their application.

The concept of a DSS is extremely broad and scientists have been researching DSSs for more than 40 years. Used in a broad range of applications, a DSS supports operations decision making, financial management, strategic decision-making in business organizations, military, logistics, etc., at different levels of an organization. In this work, we apply DSS principles to an engineering decision process.

A historical overview of DSSs is given in [24], along with a classification of DSSs into five distinct categories, summarized as follows:

- A **model-driven DSS** emphasizes access to and manipulation of financial, optimization and/or simulation models;
- A **data-driven DSS** emphasizes access to and manipulation of data, provides query tools, includes ad-hoc interpretation and visualization tools, data warehousing, etc.;

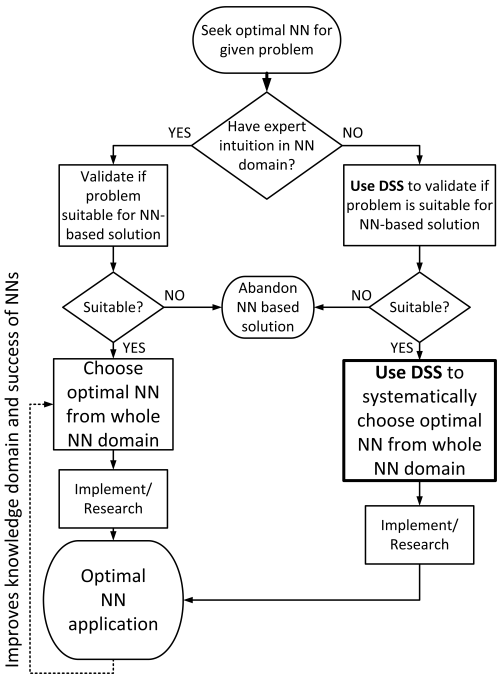


Figure 2. Our proposed DSS improves the NN design process by enabling users to systematically find optimal NN instances for their application.

- A **communication-driven DSS** uses network and communications technologies to facilitate decision-relevant collaboration and communication between decision makers and consider communication technologies as the main architectural components;
- A **document-driven DSS** uses computer storage and processing technologies to provide document retrieval and analysis, with a search engine being the core architectural component and decision-making tool;
- A **knowledge-driven DSS** can suggest or recommend optimal actions to users, based on specialized problem-solving expertise, incorporated into the DSS.

After review of DSS theory in existing literature, we decide to design our proposed DSS as a Knowledge-Driven DSS, which comprises the following components:

1. Knowledge base
2. User interface
3. Inference engine model
4. Communications component

Corresponding to the above, also shown in Figure 3, our proposed system comprises the following: NN Knowledge base (Section V), 3D visualization of data and qualitative relations (Section VI-A), inference engine and qualitative relations in data (Section VI-B), interaction with 3D environment and entry of objective parameters (Section VII), respectively. Section VII demonstrates the use of inference engine in two major use cases and presents visually the inference results.

IV. Inference engine decision space

The main and fundamental result of our work is the conceptualization and theoretical foundation for the inference engine

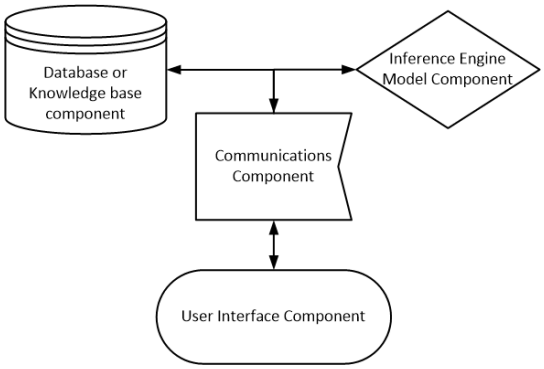


Figure 3. Components of a knowledge-driven DSS.

decision space (this section) and knowledge base framework (Section V), that enables a global overview of the complete NN domain. The decision space, presented hereby, serves also as a coherent terminology and context for our knowledge base framework. Mathematical structure of interrelations must be well defined, to facilitate an effective inference engine, used in solving multiple-objective optimization problems [25]. The decision space is defined via the following descriptors of the NN knowledge domain:

- Set of NN instances $\mathcal{I}$: contains the subset of elements of the NN knowledge domain, which are neural networks. From the whole neural network knowledge domain (NNs, research initiatives, research groups, research goals, application areas, etc.), we gather concrete NN implementations and form the set of NN instances.
- NN classifier $\zeta : \mathcal{I} \rightarrow \mathcal{P}$: provides a classification of each member of $\mathcal{I}$ into a particular set of groups $\mathcal{P}$.
- Property $\mathcal{P}$: the co-domain of a classifier ζ, with the latter considered as a function.
- Property value $p_i \in \mathcal{P}$: a specific group of some classifier. It is given a name, which is then identified as this property value.
- NN framework $\mathcal{F}$: ordered list of classifiers relevant for a given user's interest.
- NN universe $\mathcal{U}$: defined by a framework $\mathcal{F}$, it is an $|\mathcal{F}|$-dimensional space, which is Cartesian product of the properties defined by the classifiers in $\mathcal{F}$.
- NN instance $\mathcal{I}_i \in \mathcal{I}$: $\mathcal{I}_i = (p_1, \ldots, p_f)$; an f-tuple of property values, each coming from its corresponding NN property.
- NN category $\mathcal{C}_p \subseteq \mathcal{P}$: subset of a specific property, containing a set of values (classifier groups) of this property. Possibly a singleton.
- NN landscape $\mathcal{L} = C_1 \times C_2 \times C_3 \times \ldots \times C_f$ with at least one C_i being equal to the whole property P_i. Subspace of a NN universe.
- NN type $\mathcal{T} = C_1 \times C_2 \times C_3 \times \ldots \times C_f$: Cartesian product of categories. If all categories in the cartesian product are singletons, the NN type is also a NN instance.
- NN comparator δ: innate comparative quality, defining a corresponding partial order, denoted as $>_\delta$, on the set of

NN instances $\mathcal{I}$, by which some pairs of NN instances can be compared. In our proposed DSS, NN comparators are chosen by defining the NN selection criteria (see Section V). NN comparators are represented as colored arrows between NN types, with the color specifying different NN instance selection criteria and the thickness of the arrow proportional to number of evidence papers supporting the comparison.

- NN selection criteria ∇: a set of possibly competing NN comparators used for comparison of NN instances.
- Pareto front $\mathcal{R}$ of given NN selection criteria ∇: a set of (discrete) NN instances $j \in \mathcal{R}$ such that whenever some NN instance $i \in \mathcal{I}$ is better than j with respect to some NN comparator $\delta \in \nabla$, i.e., $j >_\delta i$, then there is some other comparator $\delta' \in \nabla$, such that $i >_{\delta'} j$, i.e., i is better than j w.r.t. δ'. In other words, a NN instance belongs to the Pareto front of ∇, if it cannot be improved over without harming at least one of the NN selection criteria in ∇.

What signifies our approach is the decision to abandon the aim for back-to-back comparison of specific NN implementations via rigid criteria (which would limit us to NN research subdomains) and employ a flexible DSS, enabling self-organization of data and allowing the evolution of the framework, together with the evolution of knowledge base contents.

V. Taxonomy for Knowledge Base

In contrast to related taxonomic efforts, presented in Section II, our taxonomy must provide a significant level of abstraction, allowing both a complete field overview and sufficient depth to aid qualitative comparison, while providing the flexibility for future adaptations of the proposed classification. Our generically specified ranking between feasible solutions permits us to deliver rule-of-thumb guidance that provides an excellent starting point for further in-depth analysis based on, e.g., ADT^5 criteria.

With the inference engine decision space theoretically defined in Section IV, we proceed to determine the principal dimensions for classification of NN instances. As no single source provides a definitive field overview, we as first step systematically create a taxonomic blueprint for our knowledge base. We define the NN classifiers ζ as operators for sorting of NN instances into main taxonomic branches:

ζ_1 Implementation Platform
ζ_2 NN Architecture
ζ_3 Learning Paradigm
ζ_4 Learning Algorithm
ζ_5 Learning Task

Using our defined NN classifiers, we proceed to build the taxonomy. For its core, we extract the classification used in the book Neural Networks: A Comprehensive Foundation [5], which offers a wide overview of main concepts in NN domain. To build upon this core, we add the overviews of evolutionary methods [26], Spiking Neural Networks [27] and a recent 20-years overview of hardware-friendly neural networks [28]. A principal quality of our system lies in our choice of high abstraction when defining the taxonomy; e.g., while there exist numerous flavors of the BP algorithm, our taxonomy does not differentiate between them. Only by obscuring a such detail, we can achieve a domain-wide overview. Still, as the field of NNs is very diverse, an ultimate taxonomy requires broader community collaboration and finally, consensus; both of which exceed the scope of this work.

Through processing the selected literature using our defined NN classifiers, we find that our chosen NN classifiers map NN instances into NN properties (i.e., sets of NN categories $\mathcal{C}$, possibly singletons) $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4$ and $\mathcal{P}_5$, respectively:

$\mathcal{P}_1$ (ζ_1: **Implementation Platform**) takes values from:

- General Purpose ($\mathcal{C}_1^1$): Software Simulation on general purpose computer of Von Neumann Architecture (CPU), Digital Signal Processor (DSP) Graphical Processing Unit (GPU), Supercomputer (SCP)
- Dedicated Hardware ($\mathcal{C}_2^1$): Field Programmable Gate Array, Neural Hardware / Neural Processing Unit (NPU), Analog Implementation (ANLG), Application Specific Integrated Circuit (ASIC)

$\mathcal{P}_2$ (ζ_2: **NN Architecture**) takes values from:

- Feedforward Neural Network (FFNN)
- Second Generation NNs ($\mathcal{C}_2^2$): Recurrent Neural Network (RNN),Long Short-Term Memory (LSTM)
- Spiking Neural Network (SNN)
- Cellular Neural Network (CNN)
- Self-organizing Map (SOM)
- Reservoir Networks (RSV) ($\mathcal{C}_6^2$): Echo-state Network (ESN), Liquid-state Machine (LSM)
- Convolutional NN (CONN)
- Deep Belief Network (DBN)
- Hybrid (HYB)

$\mathcal{P}_3$ (ζ_3: **Learning Paradigm**) takes values from:

- Supervised Learning (SUP)
- Reinforcement Learning (REINF)
- Unsupervised Learning (UNSUP)
- Genetic Learning (GENL)

$\mathcal{P}_4$ (ζ_4: **Learning Algorithm**) takes values from:

- Error Correction ($\mathcal{C}_1^4$, ECR): Backpropagation (BP), Extended Kalman Filter (EKF), Direct Stochastic Error Descent (DSED),
- Hebbian Learning (HBL)
- Competitive Learning (CPL)
- Evolutionary ($\mathcal{C}_4^4$, EVOL): Evolution of Architecture (EVLARCH), Evolution of Weights (EVLWT), Evolution of Learning Algorithm (EVLALG)
- Hybrid (HYB)

$\mathcal{P}_5$ (ζ_5: **Learning Task**) takes values from:

- Pattern Association ($\mathcal{C}_1^5$): Autoassociation (PASCAUT), Heteroassociation (PASCHET)
- Pattern Recognition ($\mathcal{C}_2^5$, PREC): Natural Language Processing (NLP), Principal Component Analysis (PCA), Speech (SPC), Dimensionality Reduction (DRED), Spatio-temporal (SPT)

- Control ($\mathcal{C}_3^5$, CTL): Indirect (CTLIND), Direct (CTLDIR)
- Function Approximation ($\mathcal{C}_5^5$, FAPPROX): System Identification (SYSID), Inverse System (INVSYS)
- Classification (CSF)
- Regression (RGR)

Property $\mathcal{P}_2$ thus comprises 11 NN property values, gathered in 9 categories, of which $\mathcal{C}_2^2$ and $\mathcal{C}_6^2$ each contain two property values; $\mathcal{C}_2^2$ contains p_2^2 and p_3^2 and $\mathcal{C}_6^2$ contains p_7^2 and p_8^2. Property value indices run free from category indices.

A. Qualitative comparison through NN selection criteria ∇

With the taxonomic backbone defined, we can proceed with classification of NN instances from processed literature through property values $\mathcal{P}$, using our set of NN classifiers ζ. This comparative dimension, well-defined but very permitting, is a core facility of our knowledge base and the heart of our DSS' inference engine. Therefore, we also extract from literature sources the qualitative comparison information between NN instances w.r.t. the following set of chosen NN selection criteria ∇:

δ_1 **Low cost of ownership** (feasibility, practicality, low hardware cost, low development complexity, presence of user community)

δ_2 **Capability** (effectiveness, convergence speed, generalization performance, benchmark success, high learning rate, low error)

δ_3 **Real-time requirement** (speed of execution, on-line vs. off-line learning, pre-learned vs. adaptive learning)

δ_4 **Design maturity** (proven solution vs. emerging technology)

While estimates for all NN criteria can be extracted from literature or provided by a domain expert, design maturity could also be automatically calculated as a measure of occurrence frequency in literature.

B. 5-letter notation and knowledge base formation

In the 5-dimensional NN universe, defined by our NN framework $\mathcal{F}$, that we define in Section V through selecting our set of NN classifiers $\zeta_{1,\dots 5}$, each NN instance is described via five NN properties $\mathcal{P}_{1,\dots 5}$. Therefore, each element in the database compares two NN instances or NN landscapes in terms of five parameters. To construct our formal notation, we build upon the idea of 3-letter notation used in the theory of scheduling problems [29] and adapt it to a 5-letter notation for describing NN instances. Our resulting formal representation of relation(s) between two NN instances is as follows:

$$(\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4, \mathcal{P}_5) >_{\delta_{i\dots n}} (\mathcal{P}'_1, \mathcal{P}'_2, \mathcal{P}'_3, \mathcal{P}'_4, \mathcal{P}'_5), \quad (1)$$

where each NN property $\mathcal{P}_{1\dots 5}$ can be a comma-separated list of elements (NN property values), n is the total number of selection criteria and where $i, i \in \{1 \dots n\}$ denotes the group of indices of NN qualifiers, by which the 'greater' NN instance is superior to the 'lesser' NN instance.

In our knowledge database, the following example statement extracted from a scientific source [28]: "FPGA is a superior implementation platform to ASIC in terms of flexibility and cost for implementations of FFNN or RNN with supervised or reinforcement learning, using perturbation-based error descent learning algorithms." is formally denoted as follows:

$$\begin{aligned}&(\mathcal{P}_1[4], \mathcal{P}_2[3,4], \mathcal{P}_3[1,2], \mathcal{P}_4[3], \mathcal{P}_5[x]) >_{\delta_1}\\ &(\mathcal{P}_1[6], \mathcal{P}_2[3,4], \mathcal{P}_3[1,2], \mathcal{P}_4[3], \mathcal{P}_5[x]),\end{aligned} \quad (2)$$

Or:

$$\begin{aligned}&(FPGA, \{FFNN, RNN\}, \{SUP, REINF\},\\ &DSED, x)\\ &>_{\delta_{cost,flexibility}}\\ &(ASIC, \{FFNN, RNN\}, \{SUP, REINF\},\\ &DSED, x)\end{aligned} \quad (3)$$

This example also illustrates the case where the paper does not specify all property values (in this example, the learning task $\mathcal{P}_5[x]$), the statement is incomplete and it may mean either that the relation is indifferent to that property, or that there is no information present about that property's role in the relationship. After reviewing selected literature (e.g., [28][30][26][31]–[37]), we get a number of such specific statements that comprise our knowledge base seed information, which serves as basis for development of our inference engine and visualization scheme.

VI. Result: knowledge-driven DSS with inference engine and visualization tool

The proposed inference engine, together with knowledge base visualization, are the final results of our efforts presented in this paper. Both modules operate on the data in the knowledge database in a read-only fashion. In the following subsections, we present our scheme for exploratory visualization of our multidimensional knowledge database and describe our interactive inference engine.

A. Visualization scheme

Every point in the NN universe's graphical representation corresponds to one NN instance. The most valuable information in our knowledge database is the qualitative comparison between NN instances. This is shown in Figure 4, illustrating the graphical representation of Statement (3) from Section V-B. We have found that using three dimensions for the visualization is optimal, because it allows users to navigate the environment interactively and to recognize interdependencies, even after switching between the chosen set of three dimensions. The 3D visualization can only represent three dimensions at a time and the user can explore the NN knowledge domain using any dimension set.

Figure 5 shows the 3D representation our NN universe $\mathcal{U}$, containing points from our prototype knowledge base. This view allows users to examine the NN knowledge domain in a full 3D environment, visually exploring (through zoom and rotation of view around any axis) the comparative relations between NN instances. Axes correspond to NN properties $\mathcal{P}$; each dot corresponds to a single NN instance $\mathcal{I}_i$; arrows represent qualitative comparators $\delta_{1\dots 4}$ between two NN instances; arrow thickness and dot size indicate the quantity of

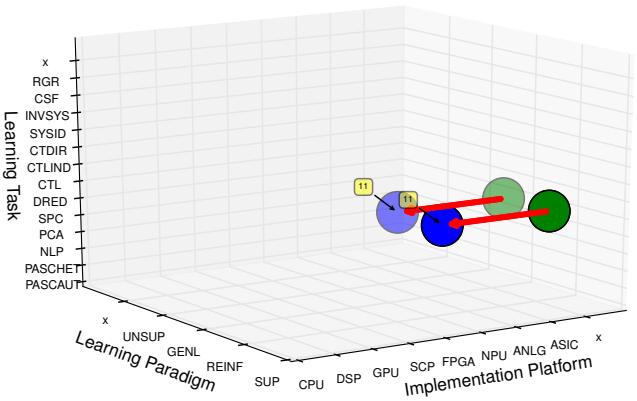


Figure 4. Example graphical representation of qualitative relation between two NN instances. The figure represents Statement (3) from Section V-B.

source papers (database entries) for the shown information; call-out-type labels are references to source literature. Each of the selection criteria is assigned its own arrow color (red, magenta, blue and black for δ_1, δ_2, δ_3 and δ_4, respectively). Coloring of NN instances aids in visual comparison (blue is better, green is worse). Using the inference engine (Section VI-B), the visualization can be actively augmented according to the user's decision input.

B. Inference Engine

Our inference engine approaches the NN instance selection process as a multiple-objective optimization problem [25] and applies a Pareto front method [38], using our disrete-equivalent Pareto front $\mathcal{R}$, as defined in Section IV, to find the suitable, multiple, non-dominant solutions. After the user specifies their boundary conditions and sets weights of the NN selection criteria ∇ through the graphical user interface, the DSS automatically identifies the discrete-equivalent of Pareto front $\mathcal{R}$ and the user can directly locate and examine the source literature, relating the NN instances in $\mathcal{R}$. Rating of alternatives is based on a weighted pairwise comparison matrix [39], resulting in levels within a discrete-space equivalent of Pareto front, which guide the user towards NN instances, specified as superior with respect to their criteria. The user can iteratively and interactively further fine-tune the selection of best candidates via weights of their criteria ∇, to determine the optimal NN instance for their problem, until the final choice is made. Information, inferred by the inference engine, is also used as input into the visualization tool, to augment the database visualization by superimposing relationships, marking Pareto points and their scores, hiding a subset of NN instances, etc. (see Figure 6). Both the visualization tool and the inference engine can be extended with additional inference and visualization functions. Section VII gives further insight into the typical application of the inference engine, through step-by-step explanation.

VII. Practical Examples of DSS use

The two major user groups that can gain remarkable benefits from using our proposed DSS, are **Industry Practitioner** and **Academic Researcher**. Both user groups share the main interest of finding the optimal NN instance for their scenario, but have a different angle: a) the industry practitioner's goal is to find the **best fitting, well proven** NN implementation for their **application** (with set boundary conditions on task type, implementation platform, etc.), and b) the academic researcher's aim is to find an active research area or synergies between domains, to systematically selects the most meaningful **research direction**.

A. DSS Use Case I: Industry Practitioner

In this section, we illustrate step-by-step a typical use case for the industry practitioner. Let our example demand a **highly accurate** and **real-time capable** NN instance for image-based **object recognition** using **supervised learning**. The following steps illustrate how this ground truth is used with our DSS as decision input and how the inference engine results are interpreted and used:

1) Enter task requirements into DSS: the practitioner enters their set of boundary conditions by selecting the NN instance properties, that are defined by the application. In our example, these are the learning task and learning paradigm. The DSS considers these two properties as pivots, therefore, only the **remaining** three properties (dimensions) are shown in the 3D visualization tool (Figure 6a). After the axes are determined, the user specifies pivot axis values (learning task = classification, CSF and learning paradigm = supervised, SUP). The effect of this is shown in Figure 6b, where only those NN instances are shown, whose pivot property values are as specified by the user. Thus, this step narrows down the search down to three dimensions and defines the NN landscape, which optimal solutions can be chosen from. If more than two pivot axes are specified by the use case, the NN landscape is 2- or 1-dimensional, further focusing the search.

2) Set weights for selection criteria ∇: once the 3D NN landscape is defined in the previous step, the user specifies weights for each of ∇ within the range from -5 to 5. In this example, the selected weights for $\delta_{1...4}$ are (see Section V-A for list of criteria):

- δ_1 **Low cost of ownership**: 2
- δ_2 **Capability**: 5
- δ_3 **Real-time requirement**: 5
- δ_4 **Design maturity**: 3

3) Examine Pareto front $\mathcal{R}$: based on weighted criteria, the inference engine extracts NN instances, that belong to the discrete Pareto front $\mathcal{R}$. These are NN instances, for which there is no NN instance superior w.r.t. any of the selection criteria (no arrows leaving the NN instance). These points are highlighted by the DSS via black squares. For better viewing of the points in $\mathcal{R}$, the user can interact with the 3D view by rotation around any axis. This is seen in 6c (left), showing the $\mathcal{R}$ points in an updated view, obtained by rotating 6b

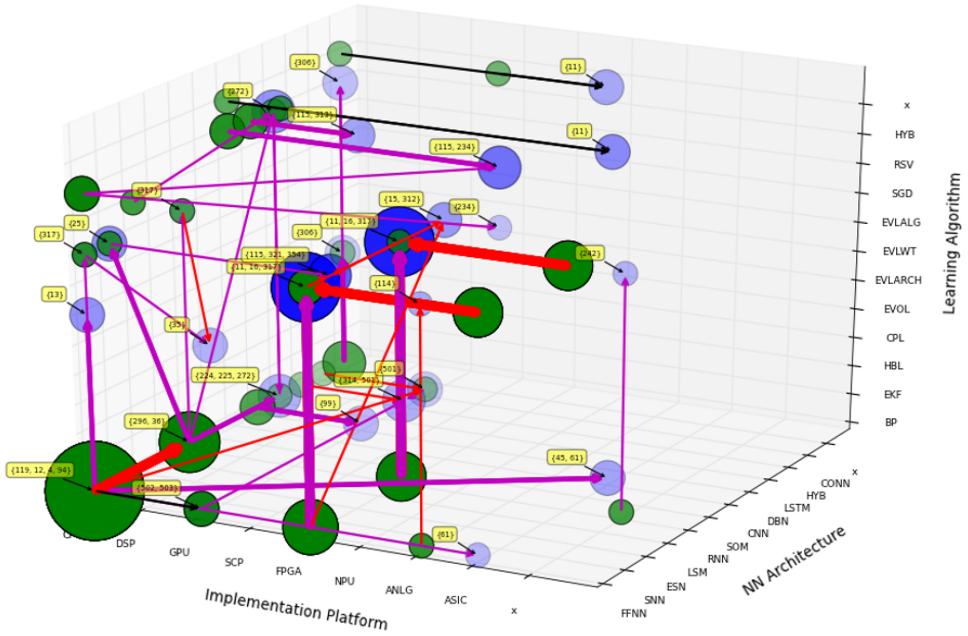


Figure 5. The 3D representation our NN universe $\mathcal{U}$, containing points from our prototype knowledge base. See Section III for axis label definitions.

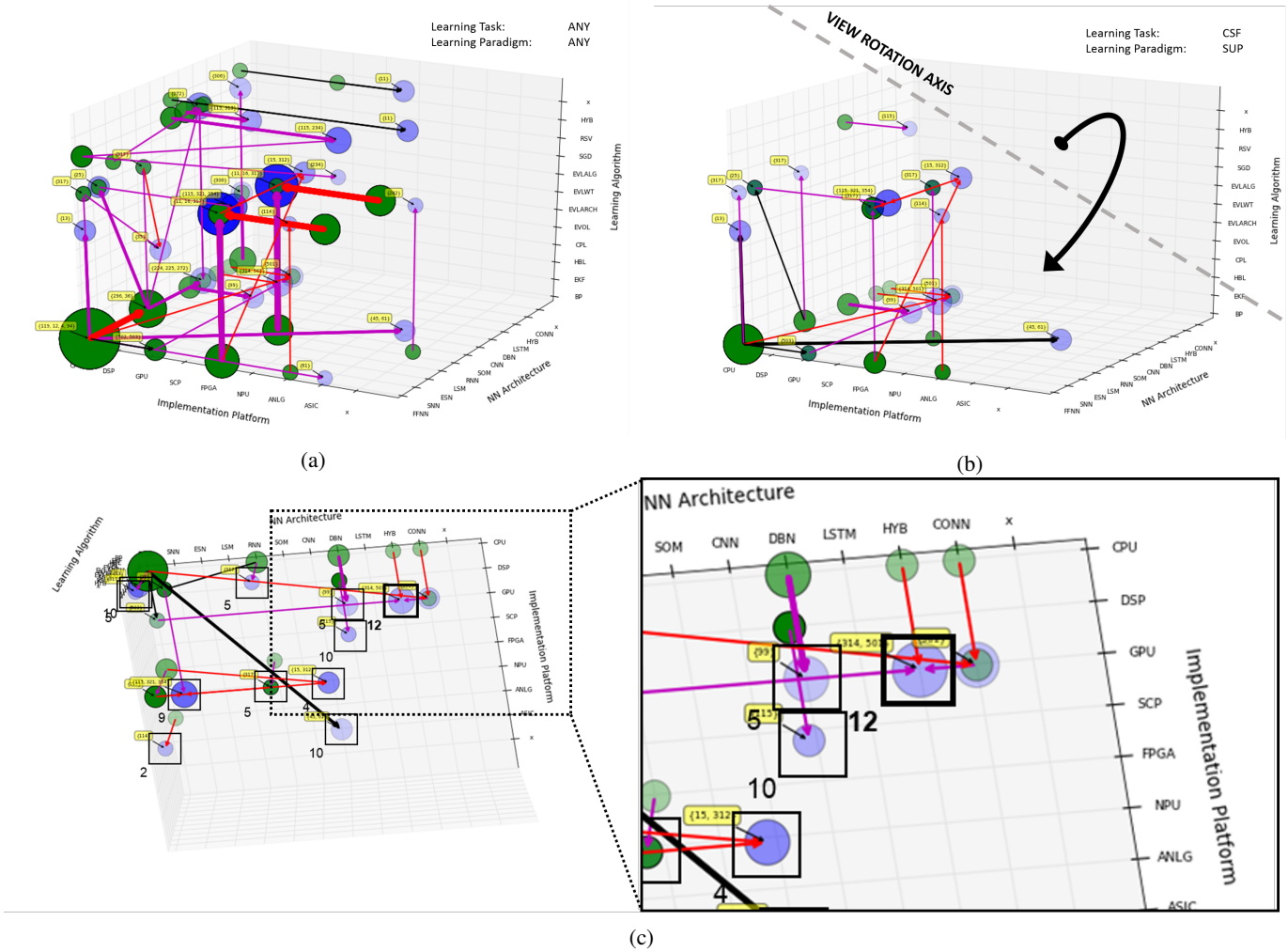


Figure 6. Rapid assessment of complete knowledge domain. Typical step-by-step application of the inference engine, together with the visualization scheme, used for finding an optimal NN instance, based on user input parameters.

around the 'view rotation axis' in the indicated direction. In the lower left corner of each $\mathcal{R}$-marking is the NN instance's score (closeup view in Figure 6c, right), calculated by the inference engine using the weighted pairwise comparison matrix.

4) Analyze top alternative in Pareto front: the user chooses the highest-ranking NN instances in the Pareto frontier and analyzes their corresponding source literature, indicated by call-outs (see Figure 4). Our example gives the highest score of 12 to NN instance, described in database entries 314 and 502 (Figure 6c). From corresponding source papers [33] and [34], the practitioner learns, that a) FFNNs can be used as convolution NNs, b) GPU implementation in [34] has better flexibility than previously known implementations, c) GPU implementation of CONN has better real-time capabilities than CPU implementation, d) Hybrid between pure CONN and FFNN has better recognition performance than any of these two used alone, e) hybrid implementation in [34] has won an impressive series of image classification competitions, etc.

In conclusion, based on the industry practitioner's input criteria, the DSS recommends, that a GPU-based hybrid CONV-FFNN NN should be investigated as best choice for the given use case. This simple case illustrates how a user can, using our DSS in a few simple steps, rapidly traverse an immensely diverse knowledge base, in order to choose an optimal direction for further investigation, and finally, concrete implementation.

5) Iterate until optimal choice is found: the user can tune the criteria to observe how the levels in the Pareto front change. Through such exploration of the NN landscape, the practitioner learns whether there exist satisfactory NN instances for their problem, given their chosen criteria, and if it exists, they systematically finds the most suitable NN instance of current state of the art for their application. Furthermore, through this exploration, the practitioner is exposed to ever new concepts and types of NNs, and at the same time also observes their relevance with respect to their own selection criteria. The database is equipped with direct references to literature sources, so the practitioner can learn about new NN concepts that directly pertain to their problem. Thus, even if there is no NN instance that fits directly with their selection criteria, the practitioner can systematically find possibilities for hybridization of seemingly unrelated approaches. This requires understanding of a broad area of NN technology and our knowledge database has a crucial role in the process. Finally, once the optimal NN instance is found or a fruitful new hybrid approach selected, the practitioner can attempt to implement the solution or conduct novel research on their own, or use our knowledge base directly to find the most relevant research groups to form collaboration. Thus, researchers can, through exploration, learn about relevant use cases for their NN types, which steers their research towards compelling yet-unknown real-world scenarios for NN use.

B. DSS Use Case II: Academic Researcher

Even though the knowledge base is constructed on existing science and lets us explore within its own, confined boundaries, it can also lead us to the discovery of new synergies, expanding the domain knowledge. It is able to do this because it can be used as a broad overview to understand the limitations of existing practices and give an indication where more research is required. This leads us to the second use case for an Academic Researcher.

For clarity of explanation of the second use case, we take the same decision input as in the case for Industry Practitioner: a highly accurate and real-time NN instance for image-based object recognition using supervised learning. However, we slightly modify the choice of weights for qualitative comparators, because an Academic Researcher is typically less interested in design maturity. Therefore, the weights used for the qualitative comparators $\delta_{1...4}$ in the selection criteria ∇ are as follows:

- δ_1 **Low cost of ownership**: 2
- δ_2 **Capability**: 5
- δ_3 **Real-time requirement**: 5
- δ_4 **Design maturity**: 1

The main significance of this use case is the notion of *pending instances* $\mathcal{I}^*$. These are *inferred* by the inference engine to direct the researcher to viable research directions, based on the highest scoring properties of existing NN instances.

The academic researcher further examines the pending NN instances $\mathcal{I}^*$ by reviewing related literature and experimenting. This is where the exploratory visualization plays a crucial role, because *through guided navigation through the NN knowledge database, the academic researcher can discover combinations of NN properties that have not yet been researched, but have high research potential.*

This gives the researcher the opportunity to critically assess the reasons why certain combinations of NN properties have not yet been tested. Furthermore, it enables the researcher to identify voids in current research and prompts them to direct their research into undiscovered domains.

The algorithm for rating and highlighting *previously-not-researched NN instances with high research potential* is:

- Step I: **Enter task requirements into DSS:** The decision input is the same as for the Industry Practitioner use case.
- Step II: **Set weights for selection criteria** ∇: The decision input is the same as for the Industry Practitioner use case.
- Step III: **Examine Pareto front** $\mathcal{R}$: In this step, points in the Pareto front are computed by the DSS to be used as input in the subsequent steps.
- Step IV: **Infer pending NN instances** ${}^*\mathcal{I}$ **with high research potential:** Using Algorithm 1, the inference engine determines the NN instances, that are not yet present in the knowledge database, but - based on the scores of NN property values of NN instances included in the Pareto front - have high research potential.
- Step V: **Visualize pending NN instances** ${}^*\mathcal{I}_i$: Superimpose pending NN instances into the 3D visualization (Figure 7a).

Algorithm 1: Infer and rank pending NN instances based on Pareto front members

Step 1: Calculate score of individual NN property values of Pareto front members:
Input: List of rated NN instances $\mathcal{I}_R \in \mathcal{R}$ according to selection criteria ∇ (based on decision input from Section VII-B)
foreach *NN property j that is not a pivot* **do**
  foreach *NN property value* p_j **do**
    foreach *NN instance* $\mathcal{I}_R \in \mathcal{R}$ **do**
      add score of $\mathcal{I}_R$ to score of p_j

Step 2: Create NN instances from all possible combinations of p_j with non-zero scores.
Step 3: From set of NN instances created in Step 2, remove all that already exist in the knowledge database. The remaining NN instances make up the set of **pending NN instances** ${}^*\mathcal{I}$.
Step 4: For each NN instance ${}^*\mathcal{I}_i \in {}^*\mathcal{I}$, compute its score through the sum of scores of its property values p (calculated in Step 1).

Step VI: **Per user specification input, limit number of displayed** ${}^*\mathcal{I}$: For better visibility, the user specifies the upper threshold for the number of pending NN instances to be displayed. In Figure 7b, this threshold is set to 30%, meaning that only the top 30% of the high-scoring pending NN instances are shown in the 3D view.

Following the steps I to VI, the researcher is drawn to areas of research, which do not yet exist in the knowledge database (see Figure 8). The researcher thus learns from the visualization, that (see Figure 9):

- in terms of NN Architecture, the top-scoring unresearched NN instances are SNN, RNN, CNN, DBN and hybrid architectures (Figure 9c)
- in terms of implementation platform, most pending NN instances are on GPU, FPGA and ASIC implementation platforms (Figure 9d)
- in terms of learning algorithm, most pending NN instances use backpropagation (BP), direct stochastic error descent (DSED) and Evolutionary learning (the variant based on evolving weights, EVLWT). (Figure 9e).
- overall, top-five best scoring pending NN instances are (Shown in Figure 8 and, with context, in Figure 9f):
  - (GPU, CNN, SUP, BP, CSF); score: 165
  - (GPU, SNN, SUP, BP, CSF); score: 164
  - (GPU, RNN, SUP, BP, CSF); score: 142
  - (FPGA, CNN, SUP, BP, CSF); score: 139
  - (FPGA, SNN, SUP, BP, CSF); score: 138

A choice of a different set of input parameters yields a completely different set of interest points and viability scores. Using different settings, the academic researcher without a narrowly-defined problem area can explore different problem areas and select their research direction based on DSS-determined viability score and their own interests, e.g., preferences regarding implementation platform, depending on previous knowledge.

VIII. Process for establishing and maintaining the knowledge database

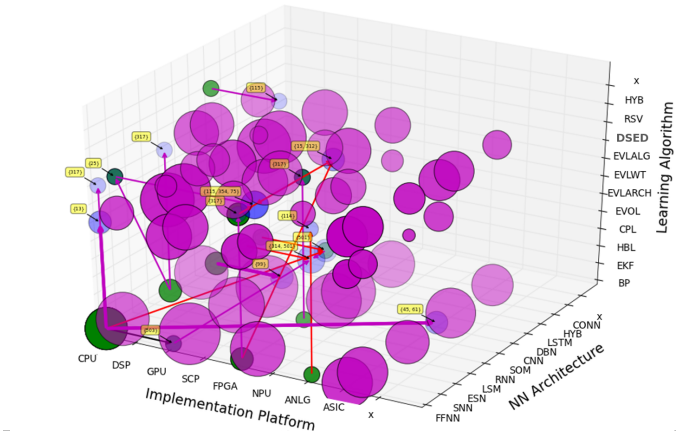
In this section, we propose concrete steps to fully exploit the potential of our results and develop our initiative further. The presented foundation, laid out as result of our research, if adopted and its potential fully exploited, can bring industrial applications of NNs closer to the forefront of today's NN research.

A central NN authority, ideally the International Neural Network Society (INNS) in coordination with IEEE Computational Intelligence Society (IEEE CIS) could consider endorsing this effort and form a working group to refine our taxonomy for use as a definitive basis for the proposed knowledge base. Once the knowledge base reaches critical mass, researchers could be motivated to contribute their own work or populate it with papers from where they notice a lack of coverage. For initial population of the knowledge base, a special issue journal could attract survey papers of NN instances and relations, represented fully in the 5-letter notation. The editorial board could, per need basis, revise the taxonomy when novel NN properties or categories emerge. When authors embrace the 5-letter notation in their papers, this information could, after the review process, be parsed and input into the knowledge base automatically.

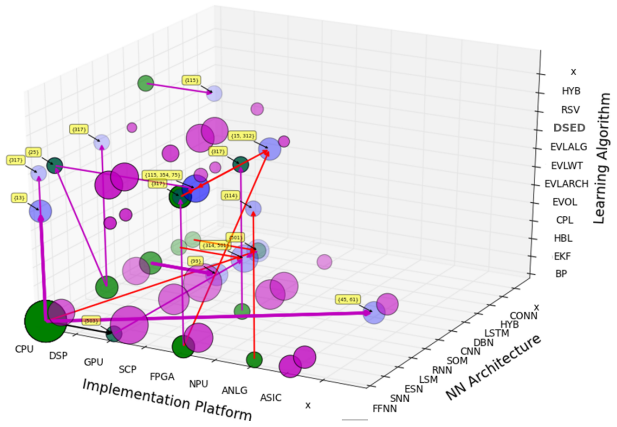
An online resource could be provided where the knowledge base, enabling navigation through the visual representation, would be freely accessible. A moderated collaborative editing among researchers could also be considered. Thus, also the expert system users themselves could submit data gathered from studying their chosen domain. An automatically-generated dynamic survey paper could be always kept up-to-date and available in printed form for quick overview of recent developments. For the database to reach critical mass, initial effort may be supported by a central NN authority. The NN community could thus set a precedence and the conceptual solution could be transferred also to other research fields.

IX. Conclusion and Future Work

In this work, we have identified the need for an abstract-level overview of the NN knowledge domain and alleviate the barriers, which an industry practitioner or researcher meet, when selecting the right NN instance or research direction for their specific scenario. We devised a theoretical foundation for a DSS, comprising a knowledge database and inference engine, that can automate the decision process of choosing the best NN architecture for the task at hand. We also presented a prototype implementation and a proof-of-concept through step-by-step use of our DSS. Next, we demonstrated an extension to the inference engine, which support automatic inference of



(a) The DSS inference engine infers pending NN instances ${}^*\mathcal{I}$, using NN property values of Pareto points inferred based on user input. Pending NN instances are marked as magenta circles, superimposed onto the 3D database view. Marker size corresponds to score.



(b) Updated view of Figure 7a, after user specifies that only the top 30% of the high-scoring pending NN instances be shown. Pending NN instances are marked as magenta circles, added to the 3D database view. Marker size corresponds to score.

Figure 7. Academic researcher use case: examination of pending NN instances.

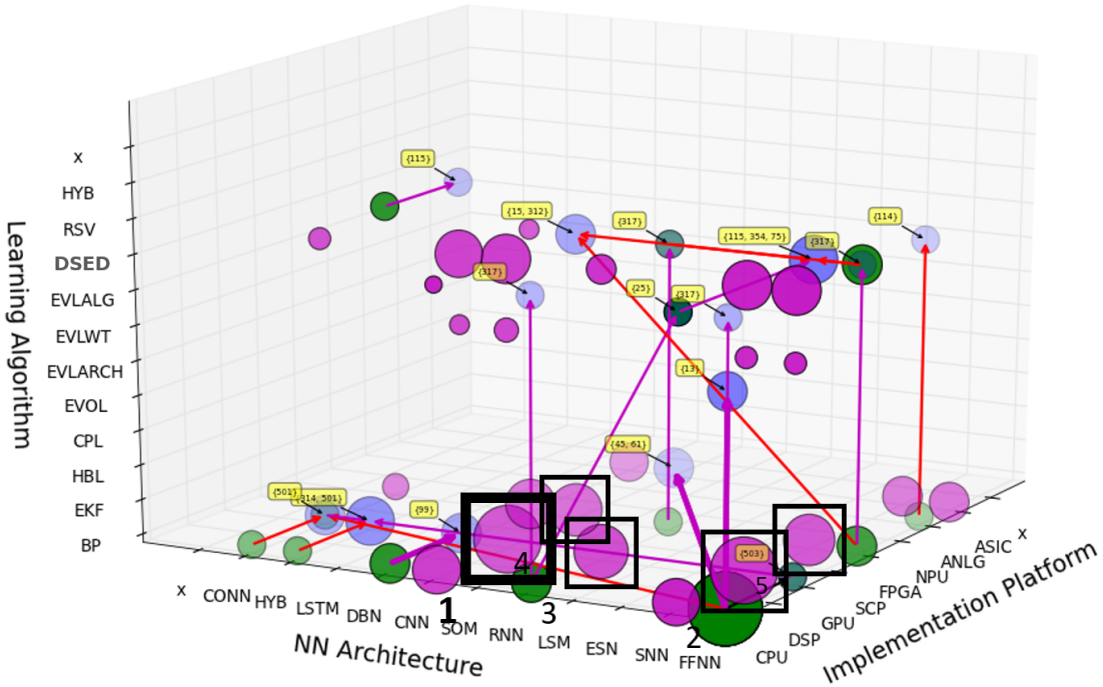
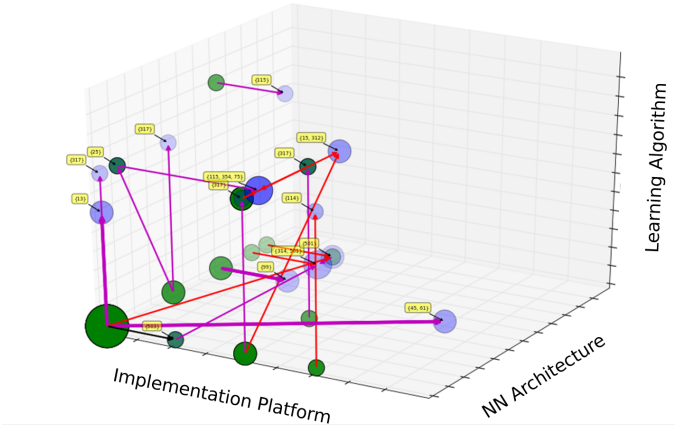
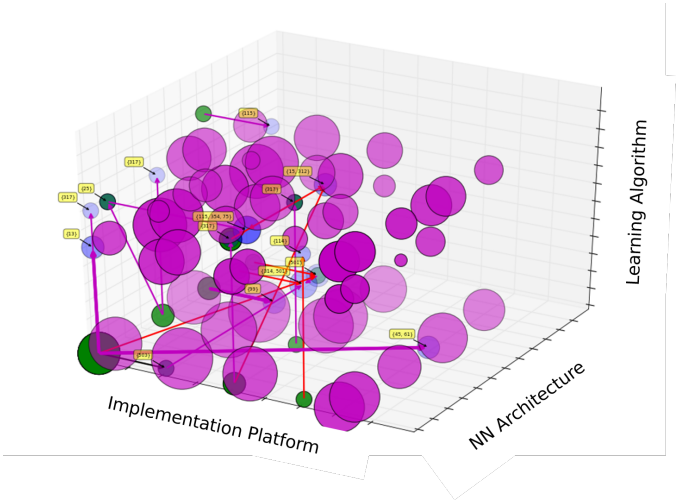


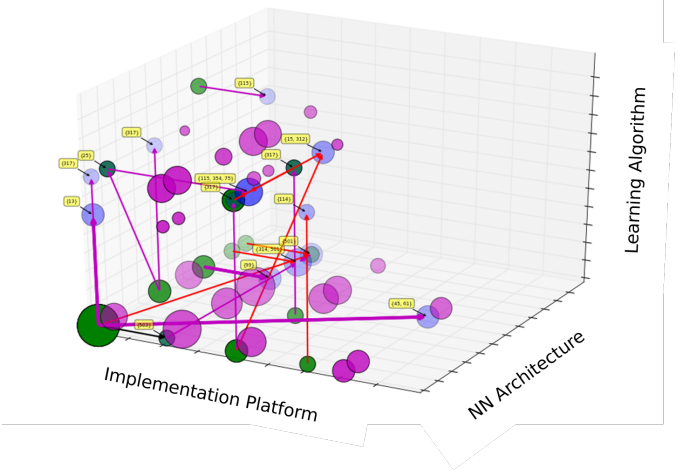
Figure 8. Best view of top-five ranking pending NN instances, marked with black squares, ranked in lower left corner.



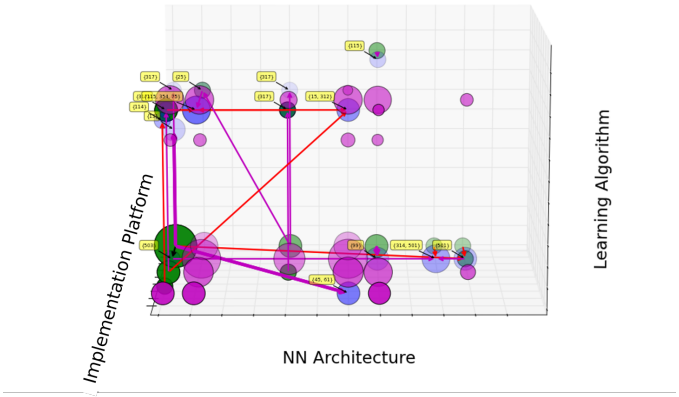
(a) NN instances that satisfy the user boundary conditions (Step I).



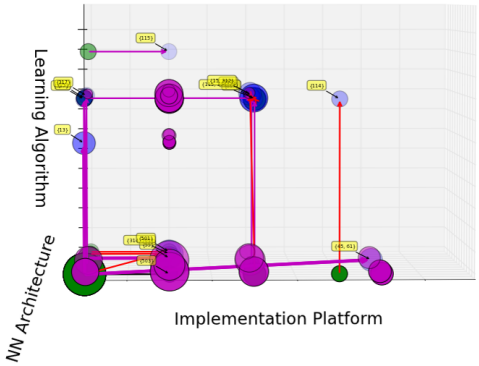
(b) Superimposed pending NN instances (Step V).



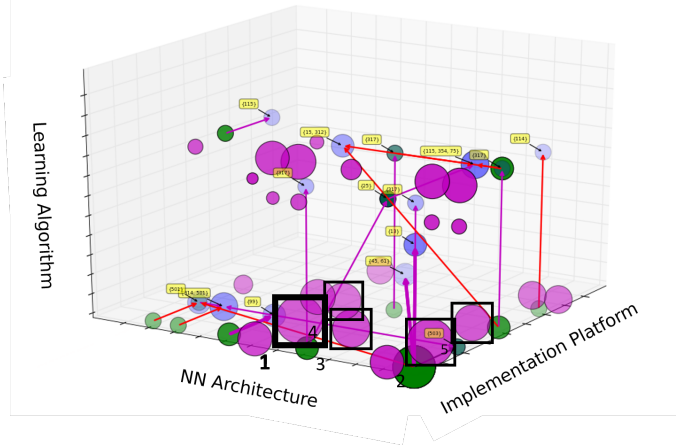
(c) Top 30% pending NN instances (Step VI).



(d) Figure 9c after clockwise rotation around Learning Algorithm axis by cca 45°.



(e) Figure 9d after clockwise rotation around Learning Algorithm axis by cca 90°.



(f) Best view of top-five ranking pending NN instances, marked with black squares, ranked in lower left corner (same as Figure 8).

Figure 9. Academic researcher use case: the DSS infers and visualizes highest-scoring pending NN instances.

promising combinations of NN properties, based on current highest-scoring NN instances within the database. This will enable our system to automatically highlight synergies between existing NN design approaches.

Future work aims towards moderated, collaborative editing of the knowledge base among researchers. The proposed 5-letter notation enables automatic parsing of the literature, keeping the knowledge database up-to-date at all times and solving this problem once and for all.

ACKNOWLEDGMENT

Operation part financed by the European Union, European Social Fund. The authors would like to thank the reviewers for their valuable comments that contributed greatly in improving this paper.

REFERENCES

[1] R. Tavčar, J. Dedič, D. Bokal, and A. Žemva, "Towards a decision support system for automated selection of optimal neural network instance for research and engineering," in Proceedings of Advanced Engineering Computing and Applications in Sciences (ADVCOMP), The Eighth International Conference on. Rome, Italy: ADVCOMP, 2014, pp. 78–85.

[2] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Networks, vol. 2, no. 5, 1989, pp. 359–366.

[3] B. M. Wilamowski, "Neural network architectures and learning algorithms," Industrial Electronics Magazine, IEEE, vol. 3, no. 4, 2009, pp. 56–63.

[4] H. Jacobsson, "Rule extraction from recurrent neural networks: A taxonomy and review," Neural Computation, vol. 17, no. 6, 2005, pp. 1223–1263.

[5] B. Taylor and J. Smith, "Validation of neural networks via taxonomic evaluation," in Methods and Procedures for the Verification and Validation of Artificial Neural Networks. Springer US, 2006, pp. 51–95.

[6] G. Dreyfus, "Modeling with neural networks: Principles and model design methodology," in Neural Networks. Springer Berlin Heidelberg, 2005, pp. 85–201.

[7] A. Omondi and J. C. Rajapakse, FPGA Implementations of Neural Networks. Springer Netherlands, 2006.

[8] Y. Huang, "Advances in artificial neural networks–methodological development and application," Algorithms, vol. 2, no. 3, 2009, pp. 973–1007.

[9] R. Rojas, "Neutral networks: A systematic introduction," Springer, 1996.

[10] C. Moraga, "Design of neural networks," in Knowledge-Based Intelligent Information and Engineering Systems. Springer, 2007, pp. 26–33.

[11] J. Ortiz-Rodríguez, M. Martínez-Blanco, and H. Vega-Carrillo, "Robust design of artificial neural networks applying the taguchi methodology and doe," in Electronics, Robotics and Automotive Mechanics Conference, 2006, vol. 2. IEEE, 2006, pp. 131–136.

[12] E. Inohira and H. Yokoi, "An optimal design method for artificial neural networks by using the design of experiments." JACIII, vol. 11, no. 6, 2007, pp. 593–599.

[13] J. F. Khaw, B. Lim, and L. E. Lim, "Optimal design of neural networks using the taguchi method," Neurocomputing, vol. 7, no. 3, 1995, pp. 225 – 245.

[14] J.-F. Qiao, Y. Zhang, and H.-g. Han, "Fast unit pruning algorithm for feedforward neural network design," Applied Mathematics and Computation, vol. 205, no. 2, 2008, pp. 622–627.

[15] H. Han and J. Qiao, "A self-organizing fuzzy neural network based on a growing-and-pruning algorithm," Fuzzy Systems, IEEE Transactions on, vol. 18, no. 6, 2010, pp. 1129–1143.

[16] S. G. Mendivil, O. Castillo, and P. Melin, "Optimization of artificial neural network architectures for time series prediction using parallel genetic algorithms," in Soft Computing for Hybrid Intelligent Systems. Springer, 2008, pp. 387–399.

[17] Z.-J. Zheng and S.-Q. Zheng, "Study on a mutation operator in evolving neural networks," Journal of Software, vol. 13, no. 4, 2002, pp. 726–731.

[18] G. G. Yen, "Multi-objective evolutionary algorithm for radial basis function neural network design," in Multi-Objective Machine Learning. Springer, 2006, pp. 221–239.

[19] M.-T. Vakil-Baghmisheh and N. Pavesic, "A fast simplified fuzzy artmap network," Neural Processing Letters, vol. 17, no. 3, 2003/06/01 2003, pp. 273–316.

[20] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," Knowledge-Based Systems, vol. 8, no. 6, 1995, pp. 373–389.

[21] A. Tickle, R. Andrews, M. Golea, and J. Diederich, "The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks," Neural Networks, IEEE Transactions on, vol. 9, no. 6, 1998, pp. 1057–1068.

[22] E. Fiesler, "Neural network classification and formalization," Computer Standards & Interfaces, vol. 16, no. 3, 1994, pp. 231–239.

[23] H. R. Maier, A. Jain, G. C. Dandy, and K. P. Sudheer, "Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions," Environmental Modelling & Software, vol. 25, no. 8, 2010, pp. 891–909.

[24] D. J. Power, Decision support systems: concepts and resources for managers. Greenwood Publishing Group, 2002.

[25] N. Xiong and M. L. Ortiz, "Principles and state-of-the-art of engineering optimization techniques," in ADVCOMP 2013, The Seventh International Conference on Advanced Engineering Computing and Applications in Sciences, 2013, pp. 36–42.

[26] S. Ding, H. Li, C. Su, J. Yu, and F. Jin, "Evolutionary artificial neural networks: a review," Artificial Intelligence Review, 2013, pp. 1–10.

[27] S. Ghosh-Dastidar and H. Adeli, "Spiking neural networks," International Journal of Neural Systems, vol. 19, no. 4, 2009, pp. 295–308.

[28] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," Neurocomputing, vol. 74, no. 1-3, 2010, pp. 239–255.

[29] R. L. Graham, E. L. Lawler, J. K. Lenstra, and A. Rinnooy Kan, "Optimization and approximation in deterministic sequencing and scheduling: a survey," Annals of Discrete Mathematics, vol. 5, 1977, pp. 287–326.

[30] L. Fortuna, P. Arena, D. Balya, and A. Zarandy, "Cellular neural networks: a paradigm for nonlinear spatio-temporal processing," Circuits and Systems Magazine, IEEE, vol. 1, no. 4, 2001, pp. 6–21.

[31] J. Schmidhuber, D. Wierstra, M. Gagliolo, and F. Gomez, "Training recurrent networks by evolino," Neural Computation, vol. 19, no. 3, 2007, pp. 757–779.

[32] G. Andrienko et al., "Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns," Computer Graphics Forum, vol. 29, no. 3, 2010, pp. 913–922.

[33] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, ""the german traffic sign recognition benchmark: a multi-class classification competition"," in Neural Networks (IJCNN), The 2011 International Joint Conference on. IEEE, 2011, pp. 1453–1460.

[34] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," Neural Networks, vol. 32, 2012, pp. 333–338.

[35] R. Raina, A. Madhavan, and A. Y. Ng, "Large-scale deep unsupervised learning using graphics processors," in ICML, vol. 9, 2009, pp. 873–880.

[36] D. Coyle, "Neural network based auto association and time-series prediction for biosignal processing in brain-computer interfaces," Computational Intelligence Magazine, IEEE, vol. 4, no. 4, 2009, pp. 47–59.

[37] R. K. Al Seyab and Y. Cao, "Nonlinear system identification for predictive control using continuous time recurrent neural networks and automatic differentiation," Journal of Process Control, vol. 18, no. 6, 2008, pp. 568–581.

[38] M. Farshbaf and M.-R. Feizi-Derakhshi, "Multi-objective optimization of graph partitioning using genetic algorithms," in Advanced Engineering Computing and Applications in Sciences, 2009. ADVCOMP '09. Third International Conference on, Oct. 2009, pp. 1–6.

[39] S. Ghodsypour and C. O'Brien, "A decision support system for supplier selection using an integrated analytic hierarchy process and linear programming," International Journal of Production Economics, vol. 56–57, 1998, pp. 199 – 212, production Economics: The Link Between Technology And Management.

Foundations of Semantic Television

Design of a Distributed and Gesture-Based Television System

Simon Bergweiler and Matthieu Deru

German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
Email: firstname.lastname@dfki.de

***Abstract*—The innovations in information and communication technologies change our daily life and the way how to interact with intelligent systems. Powerful computers are becoming smaller and are integrated almost anywhere, even in televisions. Today's connected television systems are offering a lot of technical functionalities including these, which are currently integrated in smartphones. In this article, we describe an innovative approach in form of an intelligent television system named *Swoozy*, which enables viewers to discover extended information, such as facts, images, shopping recommendations or video clips about the currently broadcast TV program by using the power of technologies of the Internet and the Semantic Web. Via a gesture-based user interface viewers will get answers to questions they may ask themselves during a movie or TV report directly on their television. These questions are very often related to the name and vita of the featured actor, the place where a scene was filmed, or purchasable books and items about the topic of the report the viewer is watching. Furthermore, a new interaction concept for TVs is proposed using semantic annotations called *Grabbables* that are displayed on top of the videos and that provide a semantic referencing between the videos' content and an ontological representation to access Semantic Web Services.**

***Index Terms*—interactive television system; Semantic Web Technologies; video annotation; gesture-based interaction.**

I. Introduction

With the growing popularity of smartphone applications (apps) a new trend slowly appeared to integrate these capabilities into television systems. In fact, the so-called connected television systems provide a wide range of technical capabilities that opens the viewers new possibilities to communicate and interact with the Internet and its services with similar features their smartphones would currently provide. This article describes an innovative approach in form of an intelligent television system named *Swoozy* [1]. This self-designed and implemented system enables viewers to discover extended information, such as facts, images, shopping recommendations or video clips about the currently broadcast TV program by using the power of technologies of the Internet and the Semantic Web.

A study conducted by the German marketer for audiovisual media SevenOneMedia [2] reveals that in a viewer panel aged between 14 and 29, 45 % of them are surfing in parallel of watching television and that the main purpose of this browsing activity is to find out more information about the program, e.g., an actor's name or biography, a location or a depicted product. This search is likely done by either using a mobile or TV app or by proactively typing in a keyword or complete phrase in a Web search engine.

The current development trend in interactive connected television systems is very app-oriented: users must install a lot of single apps, for example, one for searching videos another one for images in order to get the information they are looking for. Another technology widely spread in Europe is the Hybrid Broadcast Broadband TV standard (HbbTV) that certainly offers viewers an alternative to apps, but is currently still limited in interaction and search possibilities. These trends and technologies are described in detail in Section III.

The usage of these solutions also reveals another problem: the constant switches between several apps will oblige the user to leave his TV program and to interact several times with his remote controller before finally getting the information he was looking for.

To solve these interaction issues, the discussed approach presents a new way how viewers can interact with additional content while watching a TV program. In fact, with our solution, they are able to search in parallel for information in the Web and easily browse through the found results without an interaction breach. In its first version, the developed prototype system relies on semantic annotations gained out of the analysis of a broadcasted video combined with gesture-based interactions that will enable users to directly start a search in the Web using Semantic Web technologies, to get precise additional information in relation to the current shown scenery, like further videos, text or news articles, pictures, and furthermore shopping recommendations.

Whereas system prototypes like NoTube [3] and others [4][5][6] are using the Semantic Web for detecting possible matches between the watched program and other Web-based contents to only offer a personalized TV access, our approach uses semantic technologies on several levels. The first level is the extraction of knowledge and concepts from an ordinary non pre-annotated Digital Video Broadcasting (DVB) data stream (also called video signal). From this DVB data stream, the required information is extracted and transferred via matching rules into annotations. Over an intuitive dedicated gesture-based graphical TV interface, presented in Section V, the viewer can easily trigger a search using semantic queries. These queries are finally processed by a specially designed and implemented

backend engine called Joint Service Engine (JSE), which uses the Semantic Web, ontologies and semantic mappings to return context and domain sensitive results, as described in Section VI.

The prototype was implemented in form of set top box-based software solution to demonstrate the technical feasibility of a gesture-based interactive television system combined with semantic processing, even if the current broadcasting infrastructures do not fully provide all annotations and information required for this task. In Section II, this paper gives an overview of existing and used Semantic Web technologies and shows how annotations and semantic information can be extracted after an audiovisual analysis of the TV signal. A technical overview of currently available TV systems including the functioning of HbbTV is given in Section III. This state of the art is necessary to better delimit the core aspect of our approach from the ones, which are commercially available. Section IV presents in detail each implemented module used during the extraction process. In Section V, the choices for the design of the user interface are motivated and the method how gesture interactions lead to a semantic search is presented. Section VI will give an insight view on how the Semantic Web is used to query and deliver enriched multimedia results to the viewer.

II. Related work

A. Semantic Web technologies

The power of the Semantic Web [7] and its related technologies resides in the fact that several information sources on the Web can be used in different combinations to establish new relations between conventional semantic representations of knowledge, such as ontologies, Resource Description Framework (RDF) triple stores [8], and common Web service interfaces in form of service mashups [9].

The World Wide Web Consortium (W3C) has declared ontologies as an open standard for describing information of an application domain and also defined appropriate ontological description languages such as RDF(S) [8][10] and OWL [11]. Ontologies, as specification languages have been specially developed for a usage within the Semantic Web and mainly consists of concepts and relations. Relations organize concepts hierarchically and put them together in relationship. These relations provide a quick access to important information in a given domain, like the biography of a presenter or speaker, interesting books or shopping items. Figure 1 shows an example of how those relations can be used to find out more information about the TV program TopGear. Starting from the TV show the three main characters, Jeremy Clarkson, Richard Hammond, and James May can be found, with further references to written books or produced DVDs. A further conclusion based on all of these relations leads to a science show named Brainiac that was also presented by Richard Hammond a few years ago.

But, in order to give viewers the access to these new relations and their contents, a relation between the video's content and its semantic representation must be established: the viewed video must be annotated or more precisely a mapping

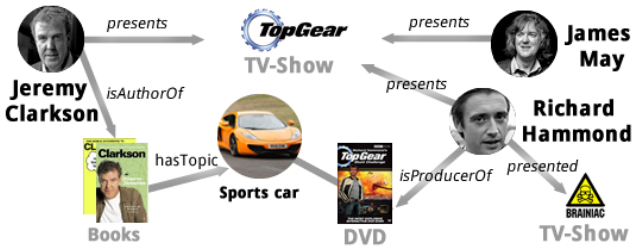


Fig. 1: Discovering new semantic relations in a TV domain.

between what the viewer is currently seeing (e.g., *a person is speaking*) and the full scene description (e.g., *this person is a politician named Barack Obama, he is the President of the United States and is giving a speech*) along with semantic annotations must be achieved through semantic mapping. This mapping combines visual information from the current scene and ontological concepts like (person, fictional character, object, and monument). Through this assignment, extracted domain knowledge can be classified [12]. This gain of knowledge out of a video can only be realized by video-based annotations: in our system we call these semantic terms or *Grabbables*.

Although several tools [13][14] and solutions exist for embedding metadata and annotations along with video - most of them are working with XML-based annotation formats like Broadcast Metadata Exchange Format (BMF) [15], Extensible Metadata Platform Format (XMP) [16], DCIM, or even MPEG-7 [17] - the core problem resides in the fact that all these metadata containing precious information are currently not transported as part of the DVB-stream, meaning that there is no possibility to reuse the semantic information of these metadata, as these are mainly used during the production workflow and not made available for further usage. Television channels certainly could provide this semantic information over an additional interface (e.g., over a Web-based REST-API access), but unfortunately this is currently not the case.

B. Semantic Web services

Semantic Web services play a central role in the presented approach as they will deliver additional contents. To achieve a correct and coherent mapping between the semantic terms and what has to be found (e.g., biography, pictures), an internal Web service ontology is needed. The latter will define how the Web services have to be accessed in term of interfaces and result types. One language to internally describe these semantic Web Services is the Web service ontology language (OWL-S) [18]. OWL-S is based on the Web Ontology Language (OWL) [19], a recommendation of the W3C, and extends its set to structures that include properties, specificity and dependencies of the Web service and express them in machine-readable and processable structures.

A concrete service description in OWL-S is divided into three parts: the *service profile*, a *service model*, and *service grounding*. Primarily the service profile is used for service discovery and describes what the service does. It contains

information about the organization that provides the service, the preconditions, input and output values, and effects, as well as the features and benefits of the service. Once a service has been selected, the *service profile* is no longer used. For the concrete process of service execution the description defined in the *service model* is used. Figure 2 shows the main concepts and relations of a service description defined in OWL-S.

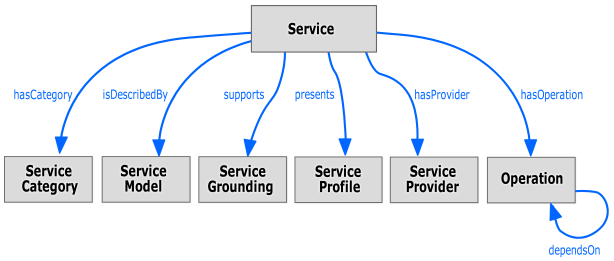


Fig. 2: Main concepts of the Web service ontology language OWL-S.

The *service model* describes the actual execution process of a service. Here, this process description consists of simple atomic processes or complex composite processes that are sometimes abstract and not executable. The process describes the individual use of the service by clients by specifying input and output data, preconditions and effects.

The *service grounding* provides detailed technical communication information on protocols and formats as well as addressing details. Furthermore, the grounding model provides a direct link or mapping between the service model and the technical service execution level. For example, implementation details like input and output messages of the *service model* are translated into corresponding elements of the service description language. The W3C recommends and specifically describes in its member submissions WSDL, but other groundings are also possible. For a better understanding of this recommendation, it is important to know that W3C member submissions serve as input to the standards process. These descriptions contain concrete information for the service implementation and realization by enabling a direct link between the grounding and the WSDL elements. In their research articles, Sirin et al. describe their prototypical implementation to directly combine OWL-S with actual executable invocations of WSDL [20][21].

The in here presented approach uses the JSE and its *Semantic Service Repository* module, described in detail in Section VI-F, that is based on the preliminary work in the area of semantic Web services modeling with grounding in WSDL and expands the approach to lightweight REST-based interfaces with their service descriptions in WADL [22][23].

C. Video annotation

Prior to any user interaction with the video stream, a processing mechanism is needed to be able to detect and analyze the actual video content. Here "analysis" describes the process of assigning a unique meaning to a video description and to be able to extract some key features such as who is presenting (name of show host, name of actor), the nature of the program (news, series, cartoon), the topic of the program ("Interview with", "News report", "Music Clip") and also objects or monuments along with their respective names and geographical coordinates.

D. Video based analysis

The first straight forward solution is to use video and visual pattern recognition algorithms to do a pixel-based analysis of each video frame as described in [24][25][26] to get the intrinsic context [27][28] of the video (e.g., a plane is landing, or a person is speaking).

Although these approaches might be suitable, they will always need training sets [29] and computational time to consolidate the results by detecting and removing false positives and to, finally, get a fully semantically annotated video frame description [30][31][32]. The prototypical implementation of *Swoozy* uses the Open CV framework to realize the video-based analysis. In order to refine the results, an additional source of information like a DVB MPEG-2 stream is needed.

E. MPEG-2 stream-based analysis

Several types of possible additional sources of information that are embedded in the MPEG-2 stream [33][34][35][36] and used in broadcast systems like DVB were identified. As specified in [36][37], the MPEG-2 stream is delivered over DVB-T and contains several encoded tables and fields enabling contextual information the television receiver is able to decode:

- Electronic Programming Guide (EPG) information - stored in the EIT table. Depending on the broadcaster, this information can be very detailed (full description of an episode including the actor's names) or very sparse: only the name of the program along with its schedule is transmitted.
- The channel's Hybrid Broadcast Broadband TV (HbbTV) endpoint URL. Usually a Web site or application URL that can be loaded and displayed by compatible television [38]. This information is extracted from the Application Information Table (AIT) [37]. The functioning of HbbTV is described in detail in Section III-B.
- Content descriptors that are transmitted usually in form of nibbles which are 4-bit content descriptors that provide a classification of the broadcasted program type (movie, drama, news, sport).
- Teletext and closed captioning information in form of pixel tables (CLUTS) or textual information.

Depending on the country and the broadcaster's allocated bandwidth on a given frequency, the amount of content present in the aforementioned tables might vary, mostly due to the packet sizes in the transmission protocol: broadcasters will logically always privilege the image quality upon transmitting non-video related contents.

The Application Information Table (AIT) contains applications and related information that can be displayed on a compatible receiver. Within its content descriptor loop, the AIT

stores pointers to HbbTV specific information (in some cases also known as Red-Button Service). In most of the cases, this pointer is an internet URL that refers to a TV-viewable Web page. By crawling this channel specific Web page additional context can be gained and extracted.

Beside the crawling and extraction of the MPEG tables, another source for our semantic extraction engine is the analysis of Closed Captioning (CC) and subtitles. Subtitles and closed captions were initially introduced for the deaf community to assist them by giving a textual transcription of a scene in form of labels placed over the video. In cases like interviews or documentaries, the closed captioning is a 1:1 transcription of the narrator's spoken text.

All the textual information and extracted context information can be processed by textual entailment [39] and Named Entity Extraction engines that will extract information and deliver semantic concepts as annotations.

F. Named Entity Recognition

Named Entity Recognition consists in extracting information out of an unstructured text. In our case, as we want to extract detailed information about a currently running TV program, it is necessary to extract it from EPG or program description. To achieve this, we rely on the Java-based implementation of the Stanford Named Entity Recognition [40] that is able to extract and label out of a text, 7 types (also named classes or concepts): Time, Location, Organization, Person, Money, Percent and Date. The result of this analysis can be delivered in XML form where each tag includes the detected concept.

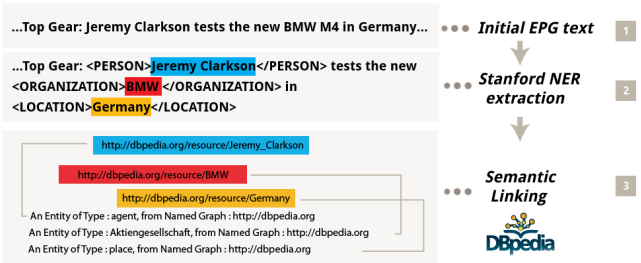


Fig. 3: Steps executed during the semantic annotation process.

Based on this first-pass analysis concepts can be linked to DBPedia [41] entities thus leading to a fully semantically annotated structure as depicted in Figure 3. This structure can then be used to define the type of *Grabbable* - a visual semantic term - to be displayed.

G. Mapping of extracted information

Once extracted from the above mentioned streams, the system classifies the extracted terms into several concepts (Person, Object, Monument, etc.), organizes them ontologically (e.g., *[Person[Politician] name: Barack Obama] [isPresidentOf] [Country, name:United States of America]*) and displays them onto the user interface in form of semantic terms. Currently our system will use a classification with following categories: Person (Actor, Politician and Speaker), Object (Car, Building), Companies and fictional Characters. Figure 4 shows how extracted streams are used to generate a visual semantic term defined as *Grabbable*.

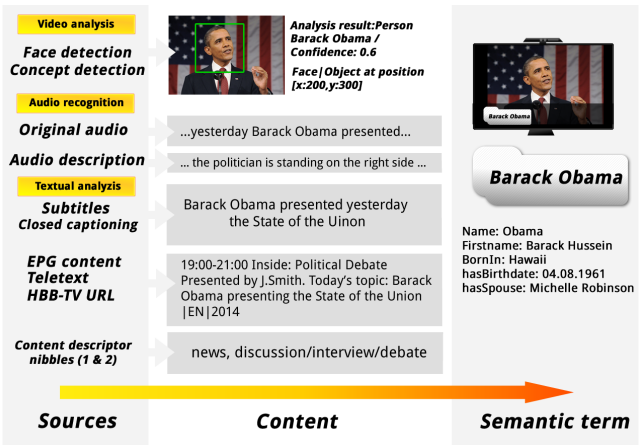


Fig. 4: Generation process of a semantic term.

H. Audio-based analysis

While the video frame-based analysis is running, an analysis of the audio channel via speech-to-text engine can be used in order to get additional details about the content. The extracted text can then be saved or delivered as a transcript and reused for an information extraction engine. In the case that the analysis of the original audio does not deliver enough information, the second possibility is to rely on the Audio Description (AD) channel. Along with the original sound of the program, an audio description provides similar to radio drama, a spoken scene description.

III. Interactive Television Systems Today: A Compact Overview

Combining television with the Internet is not a completely new idea. First systems in the late 1990's, like WebTV, later MSN TV, have shown that there is a real added value offering a unified access to email and to the Internet over a single box connected to a television device. These systems were also announcing the first wave of interaction television systems so called SmartTVs or Connected TVs that would let viewers surf and access specific services while watching their favorite TV programs. Starting from the early 2010's the market of connected TVs started to be more active and appealing through the fact that several TV manufacturers are integrating new platform technologies (e.g., Android TV, WebOS, Linux), and new interaction paradigms but also allowing third-party developers to implement their own TV-based apps.

The following section presents an overview of the state of the art in the field of interactive television systems including app-based systems. In the last part of this section, the HbbTV technology will be presented as part of an independent and broadcaster initiated approach. This overview will also focus

on design, interaction and technical implementation aspects of those platforms to better understand how the *Swoozy* approach radically differs from current commercially available systems.

A. App-based Smart TV

App-based connected TV systems are systems that are mainly focused on delivering additional content in parallel with the live television signal in form of applications or apps. This approach is borrowed from the mobile phone field, commonly known as smartphones, in which apps are very popular and playing the central interaction role between users and connected information sources. Numerous manufacturers are supporting the implementation and distribution of apps over their platforms. The following four major platforms are currently playing an essential role in the connected TV microcosm:

1) *WebOS (LG)*
WebOS is a relatively new system (2014) in the field of connected TV, but is now a well established successor of LG's first Smart TV platform called NetCast. Originally built for the now defunct Palm Pre, WebOS was licensed and brought on television by LG. Main feature of this platform is to enable users to install apps and control them over the Magic Remote - a gyroscope-based remote controller - via HTML5-based applications. The available apps are visually placed in form of icons into a bar that is present at the bottom of the television's video area. From a developer point of view, the WebOS platform allows to use third-party application programming interfaces (APIs), Javascript frameworks such as EnyoJS to build up a generic user interface and common HTML5 tools. The Software Development Kit (SDK) delivered in form of an Eclipse-based Integrated Development Environment (IDE) and an emulator helps to develop WebOS apps without having to physically install these onto the television. LG specific APIs like the Luna API are allowing a restricted access to hardware and system specific information, but unfortunately these APIs do not provide any access to EPG, channel information needed to build up a live-context centric app.

2) *Samsung TVs*
Samsung belongs to the first major manufacturers which have pushed the Smart TV concept onto the mass consumer market. Users can either use their remote controllers or hands to control a virtual cursor and to interact within apps. While using for a few years the same platform, Samsung began in the late 2014's to switch their hardware components to Tizen. The latter is an open-source operating system for numerous devices like mobile phones, wearables and TVs. This leads to the situation that currently (May 2015) two distinct SDKs are provided by Samsung: the Samsung TV SDK and the Samsung Tizen SDK. For third-party application development an Eclipse IDE framework is provided as well as an emulator to better help developers in testing their application. Over the Javascript-based Tizen Web Device API it is possible to access to additional TV channel information such as EPG or currently running show names. To support a unified visual interaction concept, a Javascript UI-Framework called *Caph* offers the possibility to easily and quickly develop UIs and apps for Tizen-based Samsung TV along with an Eclipse-based IDE for using HTML5-based UI components.

3) *Android-based TV systems*
After the early marketing difficulties of Google TV, Google has decided to persist in the television field by releasing a revamped Android-based television system. Android TV systems are currently either present in form of set-top boxes like the Nexus Player or integrated into television hardware systems sold by Philips (TP Vision). Another Android-based set-top box system is the Amazon Fire TV. All built upon Android, these systems offer viewers the possibility to interact with all apps present in the Google Play or Amazon App Store via either a remote controller, a mobile phone or even a smartwatch. The tight integration of Google-based services offers also additional features, like the online-speech recognition through the remote controller. According to the Android TV APIs it is possible to access to channel specific information like EPG although there are currently no devices on the market supporting these features. Applications for Android TV or Amazon Fire TV can be implemented in Java by using the official Android Studio IDE and the specific SDKs although other programming languages like HTML5 are also supported by the platform.

4) *SmartTV Alliance*
The SmartTV Alliance was founded in 2012 by LG and TP Vision and had as goal to setup standards and specifications for Smart TVs. Meanwhile the alliance counts over fifteen members including TV manufacturers like LG, Philips, Toshiba, Panasonic, Vestel or even IBM, which are providing white books and specifications about Smart TV Apps development. The SmartTV Alliance currently suggests to use HTML5 technologies for the implementation and to achieve this, it also provides an unified SDK for developers under the commercial motto "Build once, publish everywhere". In fact, another task of the SmartTV Alliance is to provide tooling to quickly develop apps that will then run on all the partner's hardware. The SDK includes an emulator and an Eclipse IDE for creating TV apps. A review process is then necessary before the application is accepted and dispatched to all the Smart TV alliance's devices. A full specification [42] describes each element developers should take care of during the development of HTML5-based TV apps.

5) *Initial Summary - App-based Smart TV*
Although all the platforms are providing tools and APIs

to implement apps none of them are really providing a seamless integration of both apps and live television video: in fact there is always an interaction breach between the transition from television mode to the app mode. Moreover, during our tests overlaying the current live-video with additional content was not possible. The access to live-video information like EPG or controlling the channels is very limited, on some platforms even impossible. All these elements led us to check whether HbbTV could help the viewer to access this missing information.

B. HbbTV

Hybrid Broadcast Broadband TV (HbbTV) is since 2010 a standard that specifies and defines interactive applications and additional interactive content that can be displayed by a hybrid television system [43]. It is often considered as an interactive and more appealing version of the Teletext. HbbTV was founded by the HbbTV consortium comprising several European television broadcasters (e.g., TF1, Canal+, France Television Group, ARD, ZDF). HbbTV is widely used in Europe and currently in tests in other countries like Australia or China. HbbTV also defines standardized interfaces for the usage of Internet-based technologies like IP-TV, video streaming and interactive Web-based value-added services and their technical integration into upcoming television hardware. An example of HbbTV services and the user interface is depicted in Figure 5.

Fig. 5: Example of HbbTV services (source: ARD).

Delivering technical specifications on how interactive services should be broadcasted and offered to viewers is also part of the HbbTV activities.

Hybrid means, HbbTV works in two modes. The *Broadcast* mode is controlled by the television broadcaster: he is in charge of packing into the video stream additional interactive applications. When the receiver decodes the stream, a message invites the viewer to press a button on their remote control. This is known as *Red Button* application in reference to the BBC Red Button service that since 1999 offers viewers the possibility to access additional programs and information over their television hardware. Once this button is pressed, the application is loaded (either from the video stream or from the Internet) and appears on the viewer's display.

In order to get live up-to-date information, HbbTV can rely on its second mode - *Broadband* - which will provide the link to the Internet and to the broadcaster's own online information sources (database or applications) as depicted in Figure 6.

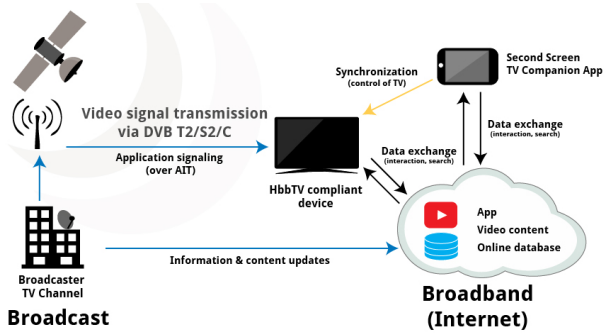


Fig. 6: HbbTV2 Architecture and technical overview.

In its second specification version, HbbTV 2 that is due in 2015, the integration of second screen and screen companions plays a major role for the broadband mode. In fact, the Internet connection is used to synchronize and push content onto the television, thus allowing a synchronization between the TV program, the information offered over the TV-based HbbTV services and the content displayed on the viewer's tablet. Although being built upon HTML-5 technologies, HbbTV contents and applications are completely working in their own ecosystem. It is not possible for third-party developers to integrate their own app or services, as only the broadcaster can decide when and which kind of services will be provided to viewers at a given moment. Moreover, from a user interaction perspective, HbbTV contents can have visually very different designs, as each channel will adopt a different layout or colors for their channel-specific applications. This leads from a user experience perspective to a lack of unified interaction and to confusion: the viewer must adapt to a new UI (including new functions and services) each time he will call the HbbTV application of a different channel.

C. Summary

Although some connected TVs rely on standard Web technologies like HTML5, the manufacturer's restrictions concerning APIs and technical aspect, do not allow third-party developers to currently leverage the full possibilities of interactive television systems. Moreover, TV-based apps running on connected TVs and the HbbTV services do not offer a sufficient and satisfiable approach to leverage the full possibilities of semantic web. Only a limited set of interaction with the content is offered to viewers, and this, within a very closed and predetermined field.

Starting from these facts and observations, the decision was made to implement a TV system that would really enable viewers to intuitively access all internet resources over user-centered interactions without any technical limitations.

IV. Architecture

The implemented system prototype is based upon a set top box plugged to a Digital Video Broadcasting Terrestrial (DVB-

T) receiver, running a customized UI, and managing interaction hardware components like a depth camera (Microsoft Kinect), a gyration mouse or a finger tracking controller (LeapMotion Controller). The functionality of these components are represented in Figure 7.

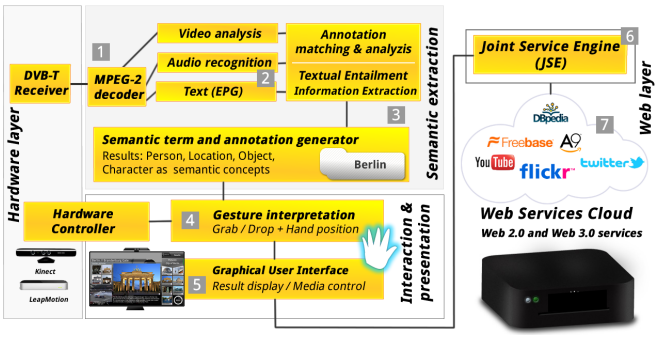


Fig. 7: Architecture of the gesture-based semantic TV system.

The architecture of the prototype system is composed of several abstract processing steps. On the one hand, there exists a user-hidden layer of signal analysis and evaluation, shown in the graphic as "Semantic extraction". This layer continuously performs an analysis of the DVB-T signal. As a result semantic terms are generated and can be used as input for a Semantic Web-based information search.

On the other hand, all user-visible processes are initiated by the user on the "Interaction and Presentation" layer. This user-centered approach gives the viewer the possibility to access additional information in parallel to the TV program by interacting with the system via a non-disruptive gesture interaction. This gesture allows to trigger a search by simply grabbing a semantic term (e.g., an actor's name). In the context of our system's approach, these terms are called *Grabbables* and must be dragged-and-dropped onto a dedicated search field element on the user interface, called *Dropzone*. This drag-and-drop interaction can be achieved, whenever the user wants to get additional information during a TV program.

Furthermore, the "Web layer" handles the connections to Web-based content. Information from different knowledge domains can be addressed via this interface, as described in detail in Section VI.

The following part lists every single processing step and task, identified by its number, presented in Figure 7. The role of the complete solution is to:

- Display the DVB-T video signal and decode the information out of the MPEG-2/MPEG-TS stream (1).
- Analyze the MPEG-2 stream and extract information out of the tables to generate corresponding annotations for the broadcasted program (2).
- Create ontologically represented semantic terms and generate graphical equivalents in form of Grabbables (3).
- Interpret gesture interactions and map them into fully formulated search queries (4).
- Use a graphical overlay principle, to enhance the user's graphical interface with additional Grabbables and multimedia annotated elements, e.g., pictures, videos, or shopping items (4-5).
- Connect via JSE to Web services, social services like Twitter, and Semantic Web Services such as Freebase or DBpedia (6-7).
- Display search results by using the interaction layer on the graphical user interface (5)

We have chosen this basis for our prototype as we are not restricted in the usage of certain APIs and have full control of both, the UI-side and the stream processing side contrary to closed proprietary solutions proposed by connected TV manufacturers.

V. User Interface and Interaction

A. Motivation for user interface design

Although aggressively promoted by current TV manufacturers, we believe that the TV app concept is not suitable for a quick search and browsing through the Web even less in the Semantic Web as described previously in Section III. Moreover, if a Web search has to be realized directly from the television set, the painfully and frustrating typing or even speaking of a keyword with a remote controller is hindering the interaction. And what happens if the viewer does not know how to spell or pronounce the name of a building in an interesting report about a city? Or the viewer does not know the name of an actor, but can recall that he was starring in an American soap? Only a long search and several switches between TV-apps and the television program might help the curious and interested knowledge hungry viewer. In some cases, this problem can rapidly turn into a decision problem, as each television broadcaster has its own app with own structures and corporate-designed interfaces leading the user to ask himself which app will be the most suitable for what he is looking for. The interaction problem is even higher when the user is zapping through several channels: must he also switch between different apps and retype his query string each time or change the context of the application manually? Unfortunately, this switching behavior brings a total interaction breach between watching the television program and getting information from the Web.

Starting from these observations, our approach tries to completely redefine the way viewers are interacting with the television by abandoning the current TV-app concept in favor of an intuitive user-centric graphical user interface.

B. User interface

The implemented graphical user interface of the created prototype system is purposely held very easy and follows all along its conception the "10 Feet Design paradigm" [44][45][46] by concentrating the efforts on having a positive trade-off between intuitive user experience, readability and easiness of interaction, so that non-computer specialists will also be able to use the system without having to cope with remote controllers and menus. Figure 8 depicts a screenshot of our current semantic television system graphical user interface.

The interface consists of a graphical overlay that will be displayed over a video: in the middle of the interface, the regular television program (e.g., received over DVB) or video stream is played. On the right, the user will find a sidebar with five thematic slots (Facts & News, Pictures, Videos, Shop, Share) that internally corresponds to specific service queries. These slots are called *"Dropzones"* and they are able to receive the created semantic terms (*"Grabbables"*). Each displayed *Grabbable* can be grabbed and dropped by the user via gesture interaction. The metaphor of the *Dropzones* is an adaption of the *Spotlets* (graphical intelligent touchscreen-based search agents) mechanism - developed in a previous Semantic Web based entertainment system [47][48][49][50].

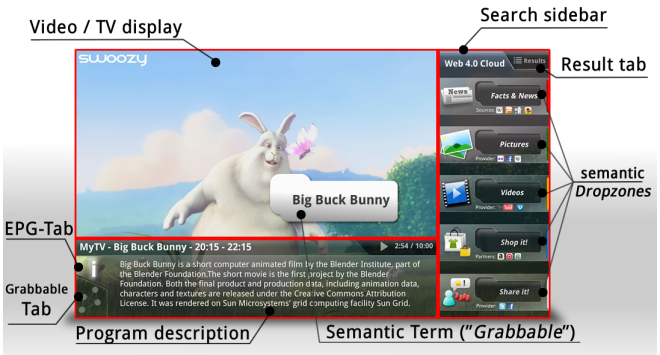


Fig. 8: Screenshot of the user interface.

The *Grabbable* dropped in one of the *Dropzones* is always annotated (Figure 9): this means a fact search about a person will have another internal meaning and output than an object search. When searching for facts about a person the search query is enriched by all extracted and represented information of the semantic description (first name, middle name, last name, gender, profession, etc.), which makes the search process of the connected JSE, described in Section VI, more effective and precise by using better filter options. For example, if the user is looking for detailed information about a building's additional properties such as the location, its architecture or inauguration date can be returned as each result has a semantic visual representation. This approach follows the "no presentation without semantic representation" paradigm [51][52][53] in usage in numerous multimodal dialog systems [48]. At the bottom of the graphical user interface, the user can either choose one of the generated Grabbables (Figure 8) or switch to the traditional Electronic Program Guide (EPG) view.

This approach breaks with the philosophy of TV apps where each app is linked to its own and single service. In this implementation, the attached JSE, is able to integrate different Web services, like Wikipedia, DBpedia, Freebase [54], Linked Movie Database (LinkedMDB) [55], Flickr or YouTube, simultaneously and it also delivers an orchestration of combined result structures. This means that the viewer will always get a unified result list, as depicted in Figure 10, where combined personal data, such as zodiac sign or portrait pictures of DBpedia and Flickr, is shown as part of the biography. In

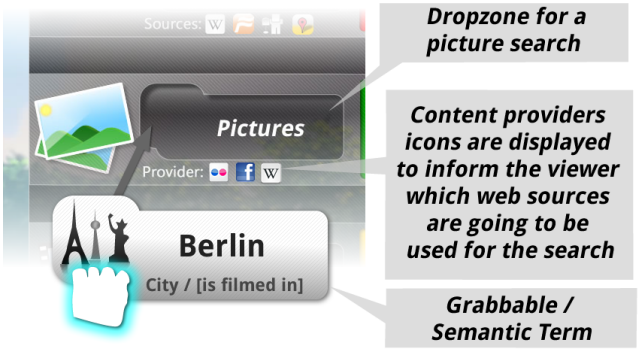


Fig. 9: Close-up of a *Dropzone*.

Figure 10, detailed facts about the famous football player "David Beckham" are displayed on the right side of the user's interface.

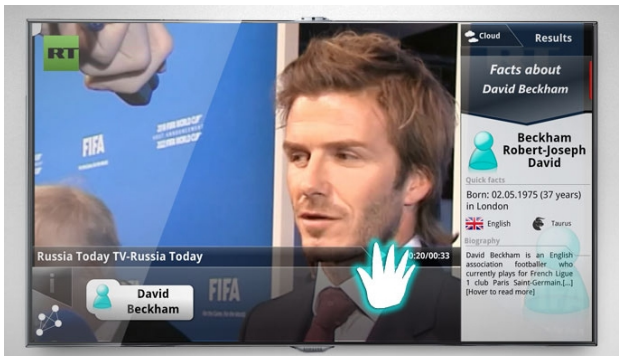


Fig. 10: Display of David Beckham's biography.

Figure 11 shows the results of a search for pictures that was triggered by a location concept named "Dubai". The pictures are retrieved from different databases and extracted by a mashup of Web services (Flickr, Wikipedia and Freebase)

Fig. 11: Picture request during a report about Dubai with results coming from different Web sources.

C. Interactions by gestures

Following the same principle of simplicity and easiness of use, we have inbuilt the possibility for the user to interact with

the system over gestures: the user only needs to move his hand towards the television screen. At this precise moment, a virtual hand is displayed (Figure 12). The position of the hand can be either tracked over a depth camera like the Microsoft Kinect, or for smaller living rooms by using a finger tracking solution, like the LeapMotion controller device [56].

Fig. 12: User gesture interaction: a virtual hand allows the user to grab out a semantic term from a video.

We have deliberately implemented only two gesture types: *Grab'n'drop* and the *Push*-gesture, as these interactions are simple to realize, do not need a specific user training and do not cause fatigue over time. The *Push* interaction is needed to make a selection and is a simplified metaphor of the traditional mouse click.

Figure 13 describes the interaction workflow. Step #1 shows how the user can grab a semantic term (*Grabbable*) from a sport report featuring Sebastian Vettel during a car show in front of the Brandenburg Gate in Berlin. In our case, the user has selected the term "Berlin" that is internally represented as a location with geo coordinates.

Fig. 13: Grab'n'drop interaction steps to trigger a picture search for the city of Berlin.

The user now would like to look for pictures of "Berlin". To achieve this, she will take the *Grabbable* (Step #2) and drop it into the Picture *Dropzone* (Step #3); within a few seconds first results coming from the Semantic Web are displayed in form of push-able elements in the right side bar (Step #4).

Beside the easiness of usage of such a system through gesture interaction, the main originality resides also in the fact that without having to type on a keyboard, or to start an additional app, any viewer will be able to rapidly get facts, videos or even shopping recommendations during his favorite TV program.

D. Mobile client application

In some cases and especially when several viewers are watching TV together, it is necessary to let them look for information without interrupting or disturbing the main television "screen". With the mobile application of our approach, depicted in Figure 14 - the mobile Swoozy App (for Android and iOS) - multiple users can simultaneously view the same TV program but interact with their own device in parallel. If viewers like to share interesting videos, pictures or facts with the other viewers, they can use the simple "sling-gesture" on their mobile device to transfer these interesting results to the TV with its large display, similarly to the 3D frisbee interaction approach presented by Becker et al. [49], where multimedia content is transferred from mobile devices to a kiosk system.

Fig. 14: Swoozy - mobile client application.

VI. Retrieval of Factual Knowledge Based on Semantic Technologies

According to the system's design, the viewer is supplied with new facts, pictures and videos while watching TV. Therefore, it is absolutely essential to access external sources to quickly find information that match exactly to the shown scenery. The presented approach uses a combination of techniques of the Semantic Web to create matching answers, while a composition of standard Web services and services of the Semantic Web is serving as knowledge source. However, the heterogeneous aspects of the services and their different Application Programming Interfaces (APIs) represent a challenge for building a correct query and coherent retrieval of matching contents. The latter must be adapted in an additional step, so that the found content can be correctly displayed onto the user's interface.

A. Motivation

As mentioned at the beginning of this article, the video, audio, and text analysis extracts knowledge concepts and adds them to predefined ontological structures, which can define persons, fictional characters, objects or locations. Via these prepared input structures out of the extraction processes, the viewers are able to trigger queries to conventional Web services or Semantic Web Services over simple gesture interaction without the need of special skills, such as programming Web service APIs or the need to learn specific database query languages like RDF(S) or query languages like SPARQL [57][58]. For non-specialists it would be very hard to formulate such queries. Indeed, these query languages are primarily used to access the full power of the Semantic Web, by allowing a navigation through semantically annotated data sets by enabling and simplifying the search for specific instances corresponding to a given request.

We assume that the typical viewer does not really want to explicitly formulate his search queries in one of the above-mentioned query languages. That is why the search will be done in the background, by using semantically annotated data sets that will be then mapped to the dropped *Grabbable*.

B. Retrieving semantic content

In order to start a search with a *Grabbable*, a dedicated engine was implemented to better solve the tasks of calling heterogeneous services and providing unified semantic results. This engine called Joint Service Engine (JSE) is involved in the retrieval of semantic content. The basic idea of the JSE is to use the joint potential of different services to focus information and knowledge. It provides and manages semantic descriptions of various pre-annotated information sources in a local *Semantic Service Repository* that opens up access to sources of different domains. This question answering component internally performs a judicious orchestration and mashing up of Web 2.0 and Semantic Web services and provides aggregated results coming from several sources - in this case Web services - as a final result. All retrieved results are returned to the client and displayed on the respective user interfaces (the television UI and/or the second screen app).

One advantage of this component is that new sources can be added, removed or replaced without hard programmatic dependencies and without stringent dependencies on specific providers of information and their interfaces. Figure 15 shows an overview of the architecture design of the specific JSE backend component adjusted for the Swoozy domain. The JSE is composed of several modules, the *Query and Presentation Manager*, the *Planning Engine*, the *Execution Engine*, the *Context Broker*, the *Mapping Core*, and the *Semantic Service Repository*, which are described in the following subsections according to the components processing workflow.

C. Query processing

The "Query" module of the *Service Engine* [59] retrieves and decomposes the user's query. The produced query structures

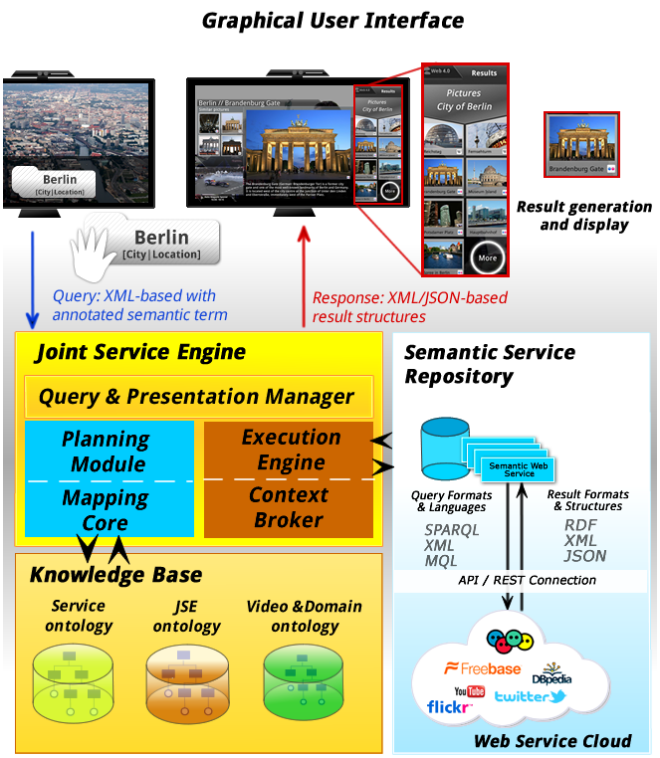


Fig. 15: Architecture of the Joint Service Engine.

are formulated according to a terminology defined by domain ontologies. They characterize the created scenery and the queries are formulated according to current generic template-based query structures, as shown in Figure 16. Each individual decomposed query part is mapped to a local meta-representation, the JSE ontology, modeled in OWL [11]. The ontology serves the purpose of precise definition of the current domain knowledge, the correct description of the retrieved content or data using a constructed reliable model based on the present situation and the environment. According to the user's query, basic ontological components like individuals are created. Based on the defined vocabulary of the JSE ontology, the planning module looks in a hard matching process for adequate plans that fulfill all of the requested properties. The resolving internal query specifies the *input type* (object, person, fictional character, company, location) specified by *properties* (complete name, keywords, etc.) and *implicit relations* and *search topics* (similar pictures, shopping facts, etc.).

One crucial point in this scenario is the discovery and execution of services. This task is executed by an execution plan which describes the discovery process by specifying which types of services are needed, what kind of domain is addressed, in which order the services have to be executed, and all the requirements needed for the matchmaking process occurring in the connected *Semantic Service Repository*. Results of the matchmaking process are ordered lists with adequately ranked information sources. The sequence of individual service

```
search topic:
generic concepts = {object (car, building),
                    person (actor, speaker, ..),
                    company,
                    location,
                    fictional character}
query-for:
  similar videos AND/OR pictures
  personal facts AND pictures AND/OR videos
  location AND/OR pictures AND/OR videos
  object facts AND pictures AND/OR videos
  shopping facts AND pictures AND videos
  sharing facts AND pictures AND videos

given properties:
  // depending on concept type
  {first name, middle name, last name, title},
  {gender, profession},
  {characterizing keywords},
  {geo-data (latitude, longitude)},
  {city-name, country-name}
  {building-name}
  {company-facts, company-name, keywords}
```

Fig. 16: Query search topics and properties.

calls which must be executed is listed in a scheduling table that needs to be processed by the *Planning Module* and the *Execution Engine*. The *Execution Engine* provides connectors and encapsulates the calls to the REST or API interfaces, by reformulating and using specific query formats like XML or languages, like SPARQL and the Metaweb Query Language (MQL). Once all results of different called services are received by the *Execution Engine*, an internal mapping process starts a review and reasoning process with the help of additional semantic mapping rules and classifies the results according to the internal JSE domain ontology.

D. Context-based brokerage

The component for context-based processing is named *Context Broker*, and is responsible for the baseline analysis of the context, prediction, and the derivation of facts. Due to its flexible and generic nature it is easily integrable into the existing workflow of the JSE. It handles the concrete extraction of the services and gathers all information and result structures. The context-dependent filtering, evaluation and explanation of data structures prevents the use of outdated or inappropriate data, and thus ensures the correct integration of knowledge structures.

The process regarding the data fusion and interpretation is executed in the following three stages:

- Merging data structures and fusion of data sources on the basis of each present context - defined by the client (*Data fusion*).
- Evaluation and transfer of information to the internal knowledge representation of the Swoozy domain - JSE Ontology (*Interpretation of data*).
- Provisioning of harmonized information that can be accessed by the *Output Presentation Manager* (*Data deployment*).

The JSE provides service functionality and access to specific data by linking community-specific data sources. As a result, this component harmonizes and consolidates the information, which is provided by several heterogeneous services like DBpedia, Freebase, Linked Movie Database or Wikidata. The *Context Broker* component distinguishes between personalization rules, mapping rules and filters. On the lowest level of complexity, there are dynamically extensible filters that check for keywords. If during a process, a filter like the blacklist filter matches, which ensures children friendly results, the data are not being taken into account at all for further processing. Figure 17 shows the *Context Broker* and the multi-stage process of filtering that forms the essential task of this component. The functionality of filter-based rules and simple filter routines is used in different processing stages of the JSE: integration, interpretation, mapping and fusion of data. The goal is, to be able to do an absolutely coherent and correct mapping of the retrieved contents to embed heterogeneous external data structures into the existing knowledge structure. A comprehensive and reliable retrieval and detection of information creates the added-value and improves the Swoozy approach.

Personalization also plays an important role within this component: based on the context description, individual context models can be defined. A rule for context modeling consists of an explicit ontological description to interpret and structure the present situation. The *mapping rules* are necessary in order to map the available contents to specific instances in the style of the internal ontology. First, for the analysis of the obtained data with the designated filters, the contents are bound in interim meta-instances and released as suitable concrete instances for the downstream modules at the end of the processing chain. All retrieved and fused data structures from the *Execution Engine* are assigned and mapped to meta-instances of the JSE-ontology by predefined patterns of mapping rules and added to a temporary list. Each mapping rule contains implicit expert knowledge, enriched and expanded with the user's personal information, which defines, how to map the connected external data sources to the internal representation, and how to solve potential conflicts between different application domains. Depending on the query, it might be necessary, to merge the meta-instances of the temporary list or to directly create links between the present meta-instances. Within the preceding orchestration process one result - for example the location where an actor was born - can be used as input for the next service, to get extended location information like longitude and latitude of the place of birth. Before the results of the temporary list can be processed internally, they are subjected to a reconsideration by special filter-based rules. This processing step is needed to filter out duplicates, to merge individual properties and to apply information extraction methods. Information extraction means the retrieval of structured information from the present texts by methods of the field of Natural Language Processing (NLP), e.g., Named Entity Recognition is used for the extraction of

facts and automated tagging of contents with topics like names of cities, places or locations, names of persons or buildings [40][60], as previously described in Section II-F. In this final step, the remaining instances from the temporary list are provided for further processing and the context broker will set up and draw new relationships between the retrieved extracted facts of the instances. This is done based on predefined patterns for persons, monuments, objects and fictive characters, e.g., *person X isBornIn city Y*.

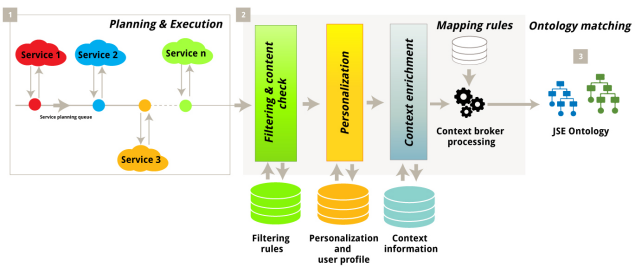


Fig. 17: Context Broker component, filtering and mapping mechanisms.

Now, after all the filters and upstream modules were introduced a concrete example will give a better overview of the internal process workflow of the module. The internal process workflow, illustrated in detail in Figure 18, is controlled by the central planning module and begins with an input structure that is formulated as a user query, e.g., “Show me more pictures of an actor!”. After the query processing the matchmaking process looks for adequate services for the decomposed query parts. As a result all present services are listed, even orchestrated services can be included. Each service gets called by the execution engine and in a preprocessing step, the incoming result structures are screened for several keywords and string characters, defined in an extendable list of multiple rules. In cases of failures (e.g., when the service is not reachable) the component will monitor it and will reschedule the search task by creating a new plan using alternate services. Content that passed the initial filtering is mapped to instances based on the internal ontology and on predefined mapping rules that exist for each service. These mapping structures form the so-called external expertise to trigger the mapping of the current data to the local domain. In an advanced filtering process, the instances are analyzed to find duplicates: the results are then weighted based on named entity recognition, personalization rules, and in a final step relations are drawn between the resulting items. All remaining harmonized result structures are linked to concrete instances that are then forming the ontological final result representation.

In the deployment stage, the *Context Broker* component prepares the internal OWL-Structure for output presentation. Customer or user’s personal information and the demanded output format or view of the output structures are deposited in the JSE-Ontology. It is then passed as is to the *Presentation Manager* to create specific output structures in a REST-client friendly format, like XML or JSON.

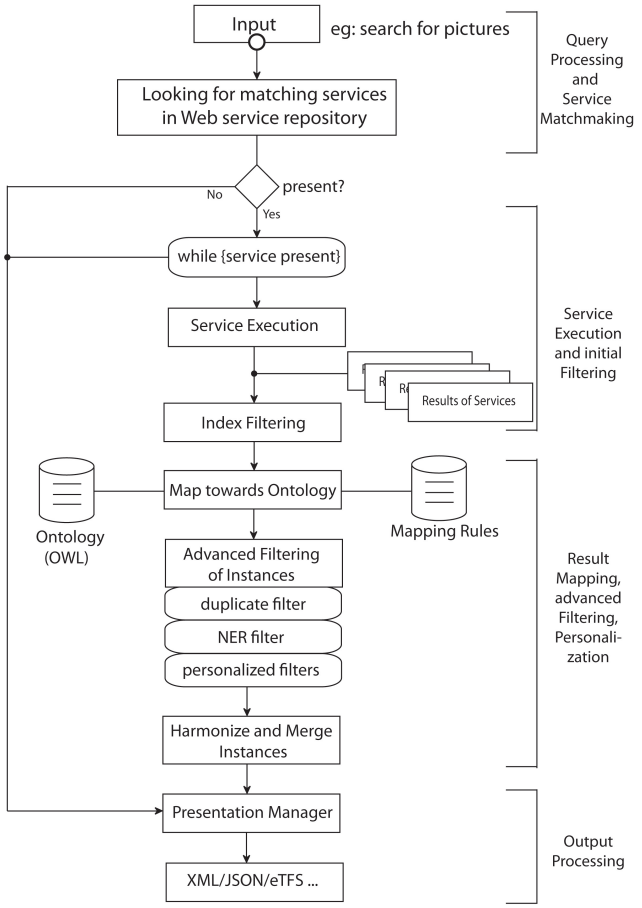


Fig. 18: Context Broker processing workflow.

E. Mapping and matching

During the processing chain inside the JSE, the content of the described data channels must be repeatedly transformed from one data format to another. The most important step for creating comparable and interoperable data models is the definition of mapping functions between the used concepts. Therefore, the identified data structures must be mapped based on stored mappings that have been defined in a pre-processing phase in a formal description language. For an unambiguous assignment of the models and types of a described element, the mapping functions are specified by categories. The *Mapping Core* achieves these mappings in each component of the JSE. In the “Query” module, the user’s query input description is mapped to the internal domain ontology that is used for further processing during the planning process. Additionally, in the “Execution” module, a mapping of the results of the called external sources to the internal JSE ontology must be fulfilled. Moreover, the range of result formats varies from simple JSON structures to complex semantic data structures like RDF. In this specific case, formal mapping rules are used to allow a higher quality data type mapping on a more generic level: new instances can be created and linked to each other. Alternatively, a taxonomy of objects can be mapped according

to their internal data structures.

F. Semantic Service Repository

The *Semantic Service Repository* provides access to different types of information sources like Semantic Web Services that cover information stored in external database management systems or Semantic Repositories. In a prototypical implementation [59] concepts for detailed service descriptions in OWL-S [18] are created and deployed in this *Semantic Service Repository*. Concrete service descriptions are realized for freely available knowledge sources, such as DBpedia, Freebase, Flickr or Linked Movie Databases. In these Web-based systems, information is stored and made accessible in a structured and manageable form, which would be otherwise difficult to access through special query languages like SPARQL for DBpedia or MQL in the case of Freebase. The main difference of this approach, compared to conventional database management systems, is the usage of ontologies as a technology to harmonize and store semantically structured data: each concept defines and classifies information and also adds implicit knowledge characterized by its name and position in a hierarchy or taxonomy [61]. The JSE also closes the gap between pure RESTful service calls and factual knowledge extracted from Semantic Web Services like Freebase or DBpedia, by mapping results and their respective annotations syntactically and semantically according to a well-defined domain ontology.

The major part on the technical side of the *Semantic Service Repository* is the discovery and matchmaking of services. The repository hosts, provides and manages descriptions of various pre-annotated sources based on Semantic Web technologies. This component provides more flexibility through simple modifications, like an add-, delete- and replace-functionality of stored service descriptions, and uses a modular design for individual components without stringent programming dependencies. Inspired by the approaches of Sirin [21] and Lambert and Domingue [23], all Web services in our *Semantic Service Repository* are represented in OWL-S with a grounding declaration in WSDL [62] or WADL [63].

The operation of the internal discovery process of the *Semantic Service Repository* was described in detail in the description of the generic flexible framework [61]. Figure 19 presents a simplified overview of the internal processing steps. A *Broker* consists of a *Query Handler* that interprets a query from the *Planning and Execution* module. The query is formulated in XML based on a single XML schema. From this interpretation, facts derive abstract types of services. By means of a *Rule Engine*, SPARQL queries can be derived from the decomposed query facts, which are used for the matchmaking process. This query is then forwarded to the *Matcher*, that connects to the service repository and executes the received SPARQL query, which points to adequate and concrete Semantic Web Service representations of deposited services. If the query matches, a list of semantically described services is returned. The used service ontology describes how to interpret and execute the external service. All required parameters for the concrete execution of the service and its external call are stored in this ontological description. The orchestration of services, their specific requirements and dependencies are also stored within this ontological description. In the latter, the grounding description of the external services is based on the technical concepts of WADL and WSDL. These are describing the technical aspects of communication and have been extended within our system with parameters and properties that refer to pattern generated SPARQL queries. Each SPARQL query also contains parameters used within the output structures. Those parameters have a direct mapping to the internal ontology and they are mainly used to call specific knowledge databases like DBpedia or triple stores.

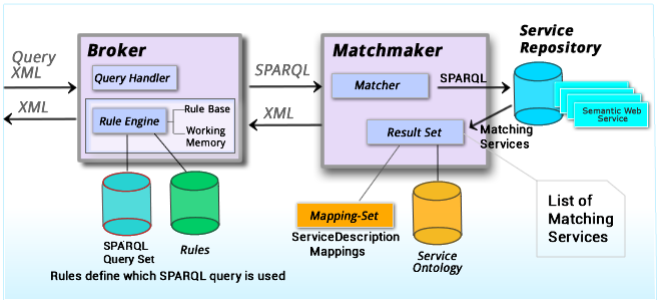


Fig. 19: Service discovery process.

A combined result structure is passed to the *Planning and Execution* module, which triggers the process of integration, analysis and harmonization of external data structures.

G. Output presentation

The last step of the processing is done by the *Presentation Manager* which will encapsulate and transform the semantic annotations in a standardized result structure. The contents of the delivered result structures are displayed on the graphical user interface (television screen) after a parsing process. Depending on the user's query, e.g., a media search, different structured output formats (RDF, XML, JSON, etc.) might be served by the *Presentation Manager* module. This module uses filter rules and generic declarative element-based mapping techniques to create the resulting structures from the internal domain ontology and returns these structures to connected client platforms. This procedure allows a parallel distributed output: both second screens and television systems are fed with the results coming from the *Presentation Manager*. With this parallel output processing a cross-media interaction is possible.

H. Data and semantic service management

The choice of which Web service is going to be internally triggered is based on rules and filters as defined in Section VI-E . These rules are not static and can be changed over time by a dedicated Web-based management board: there it is possible to adapt the queries and edit the underlying ontologies and rules used within the query process.

Moreover, via this management module, the ontology can be modified and defined mapping rules can be easily adapted: new

Web services and knowledge databases can be deployed, or activated. This aspect also leverages the modular and distributed cloud-based concept of our system: each service is easily interchangeable due to their generic ontology-based definition. The possibilities to edit rules and to select the required Web services are motivated by the pure semantic orientated approach of our system. Additionally, in parallel to free accessible internet based content, like DBPedia, it is possible to adapt the system in accordance to the broadcaster's need, to promote and give their viewers access to own premium content services (specific videos, interviews, infographics). Over the Web-based management board broadcasters can easily and quickly adapt the offered and displayed contents within the Swoozy UIs.

VII. Conclusion and future work

We demonstrate with our approach - the *Swoozy* system - and its prototypical implementation, that it is possible to provide a novel way to interact with video contents without interaction breaches. Through a seamless combination of gesture-based interaction, video information coming directly from the broadcasted signal, and the Joint Service Engine as backend service connector it is possible to enhance the displayed content with additional information from the Internet (Web or Semantic Web) and its related services. The scenery is underpinned by additional information of external services, wherein this information is context-based, machine-readable and interpretable.

An extended version of the Swoozy system is currently in usage in several living labs in Germany and connected to third-party multimedia platforms targeted to various application domains: the system has also been presented to a wider audience during international exhibition fairs, e.g., at the CeBIT, and there it generated a lot of positive user feedback, especially the innovative exploration in connected knowledge databases.

Moreover, thanks to this innovative approach, television enters into a new dimension in which viewers will receive additional information and knowledge about the persons, locations, and objects featured in their favorite television programs.

The *Swoozy* concept offers a wide range of capabilities and possibilities to communicate and is not only applicable for the sole field of television. The concept can also be used to enhance the functionality of existing video-based systems, such as video-on-demand platforms, interactive e-Learning systems, video casts or even online university courses, where the semantic terms would be mathematical formulas or technical concepts.

We believe that the concept of semantic television will turn television into an appealing and ludic knowledge provider and will give a brand new dimension to interactive connected television systems in the future. Moreover, in addition to the input modalities (Microsoft Kinect and LeapMotion controller) used in *Swoozy*, we consider extending our gesture-based approach to Smartwatches.

References

[1] M. Deru and S. Bergweiler, "Swoozy - An Innovative Design of a Distributed and Gesture-based Semantic Television System," in Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2014), International Academy, Research, and Industry Association (IARIA). IARIA, 8 2014, pp. 131–139, best Paper Award.

[2] SevenOneMedia, "HbbTV macht TV clickbar," 2013.

[3] L. Aroyo, L. Nixon, and L. Miller, "NoTube: the television experience enhanced by online social and semantic data," in Consumer Electronics-Berlin (ICCE-Berlin), 2011 IEEE International Conference. IEEE, 2011, pp. 269–273.

[4] Y. B. Fernandez, J. J. Pazos Arias, M. L. Nores, A. G. Solla, and M. R. Cabrer, "AVATAR: an improved solution for personalized TV based on semantic inference," Consumer Electronics, IEEE Transactions on, vol. 52, no. 1, 2006, pp. 223–231.

[5] J. Kim and S. Kang, "An ontology-based personalized target advertisement system on interactive TV," Multimedia Tools and Applications, vol. 64, no. 3, 2013, pp. 517–534.

[6] B. Makni, S. Dietze, and J. Domingue, "Towards semantic TV services a hybrid Semantic Web Services approach," 2010, [Retrieved: May 2015].

[7] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, 2001, [retrieved: July 2014]. [Online]. Available: http://www.jeckle.de/files/tblSW.pdf

[8] G. Klyne and J. J. Carroll, "Resource Description Framework (RDF): Concepts and Abstract Syntax," W3C Recommendation, 2004. [Online]. Available: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[9] P. Hitzler, M. Krötzsch, S. Rudolph, and Y. Sure, Semantic Web: Grundlagen [Basics]. Springer Berlin Heidelberg, 2008.

[10] R. Cyganiak, D. Wood, and M. Lanthaler, "RDF 1.1 Concepts and Abstract Syntax," W3C Recommendation, 2004, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

[11] P. F. Patel-Schneider, P. Hayes, and I. Horrocks, "OWL Web Ontology Language Semantics and Abstract Syntax," W3C Recommendation, 2004, [retrieved: May 2015]. [Online]. Available: http://www.w3.org/TR/2004/REC-owl-semantics-20040210/

[12] M. C. Surez-Figueroa, G. A. Atemezing, and O. Corcho, "The landscape of multimedia ontologies in the last decade," Multimedia Tools and Applications, vol. 62, no. 2, 2013, pp. 377–399.

[13] M. Lux, W. Klieber, and M. Granitzer, "Caliph and Emir: semantics in multimedia retrieval and annotation," in Proceedings of the 19th International CODATA Conference. Citeseer, 2004, pp. 64–75.

[14] M. Lux and M. Granitzer, "Retrieval of MPEG-7 based Semantic Descriptions," in In Proceedings of BTW-Workshop WebDB Meets IR, 2004.

[15] Institut für Rundfunktechnik, "Broadcast Metadata Exchange Format," BMF 2.0, 2012, [retrieved: May 2015]. [Online]. Available: http://bmf.irt.de/

[16] Adobe, "XMP - Adding intelligence to media," 2012, [retrieved: July 2014]. [Online]. Available: http://www.adobe.com/devnet/xmp.html

[17] J. Martnez, R. Koenen, and F. Pereira, "MPEG-7: the generic multimedia content description standard - part 1," Multimedia, IEEE, vol. 9, no. 2, 2002, pp. 78–87.

[18] D. Martin et al., "OWL-S: Semantic Markup for Web Services," W3C Submission, 2004, [retrieved: May 2015]. [Online]. Available: http://www.w3.org/Submission/2004/SUBM-OWL-S-20041122/

[19] D. L. McGuinness and F. van Harmelen, "OWL Web Ontology Language Overview," W3C Recommendation, 2004, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/owl-features

[20] E. Sirin, B. Parsia, and J. Hendler, "Filtering and Selecting Semantic Web services with Interactive Composition techniques," IEEE Intelligent Systems, vol. 19, no. 4, 2004, pp. 42–49.

[21] E. Sirin, "Combining Description Logic Reasoning with AI Planning for Composition of Web Services," dissertation, pp. 1–239, 2006.

[22] O. F. F. Filho and M. A. G. V. Ferreira, "Semantic Web Services: A RESTful Approach," in IADIS International Conference WWW/Internet. IADIS, 2009, pp. 169–180, [retrieved: May 2015]. [Online]. Available: http://fullsemanticweb.com/paper/ICWI.pdf

[23] D. Lambert and J. Domingue, "Grounding semantic web services with rules," in Proceedings of the 5th Workshop on Semantic Web Applications and Perspectives (SWAP), ser. CEUR Workshop Proceedings, A. Gangemi, J. Keizer, V. Presutti, and H. Stoermer,

Eds., vol. 426. CEUR-WS.org, 2008, [retrieved: May 2015]. [Online]. Available: http://ceur-ws.org/Vol-426/swap2008_submission_8.pdf

[24] S. Bloehdorn et al., "Semantic Annotation of Images and Videos for Multimedia Analysis," in The Semantic Web: Research and Applications, ser. Lecture Notes in Computer Science, A. Gómez-Pérez and J. Euzenat, Eds. Springer Berlin Heidelberg, 2005, vol. 3532, pp. 592–607.

[25] E. Sgarbi and D. L. Borges, "Structure in soccer videos: detecting and classifying highlights for automatic summarization," in Progress in Pattern Recognition, Image Analysis and Applications. Springer, 2005, pp. 691–700.

[26] W. Shao, G. Naghdy, and S. Phung, "Automatic Image Annotation for Semantic Image Retrieval," in Advances in Visual Information Systems, ser. Lecture Notes in Computer Science, G. Qiu, C. Leung, X. Xue, and R. Laurini, Eds. Springer Berlin Heidelberg, 2007, vol. 4781, pp. 369–378.

[27] L. Ballan, M. Bertini, and G. Serra, "Video Annotation and Retrieval Using Ontologies and Rule Learning," IEEE MultiMedia, vol. 17, no. 4, 2010, pp. 80–88.

[28] U. Arslan, M. E. Dönderler, E. Saykol, Ö. Ulusoy, and U. Güdükbay, "A Semi-Automatic Semantic Annotation Tool for Video Databases," in Proc. of the Workshop on Multimedia Semantics (SOFSEM 2002), ser. SOFSEM-2002, 2002, pp. 1–10.

[29] G. Quénot, "TRECVID 2013 Semantic Indexing Task," 2013.

[30] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, no. 9, 2010, pp. 1627–1645.

[31] C. Snoek, D. Fontijne, Z. Z. Li, K. van de Sande, and A. Smeulders, "Deep Nets for Detecting, Combining, and Localizing Concepts in Video," 2013.

[32] L. J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," in Advances in neural information processing systems, 2010, pp. 1378–1386.

[33] T. Dajda, M. Cislak, G. Heldak, and P. Pacyna, "Design and implementation of the electronic programme guide for the MPEG-2 based DVB System," 1996.

[34] C. Peng and P. Vuorimaa, "Decoding of DVB Digital Television Subtitles," Applied Informatics Proceedings - No.3, 2002, pp. 143–148.

[35] M. Dowman, V. Tablan, H. Cunningham, C. Ursu, and B. Popov, "Semantically enhanced television news through web and video integration," in Second European Semantic Web Conference (ESWC'2005). Citeseer, 2005.

[36] "Digital Video Broadcasting (DVB) Subtitling systems," European Standard ETSI EN 300 743, European Broadcasting Union, 2014.

[37] "Specification for Service Information (SI) in DVB systems," European Standard ETSI EN 300 468, European Broadcasting Union, 2014.

[38] K. Merkel, "HbbTV - Status und Ausblick," 2012, [retrieved: May 2015]. [Online]. Available: http://www.irt.de/webarchiv/showdoc.php?z=NTgwNyMxMDA1MjE1I3BkZg==

[39] R. Wang and G. Neumann, "Recognizing Textual Entailment Using Sentence Similarity Based on Dependency Tree Skeletons," in Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, ser. RTE 07. Association for Computational Linguistics, 2007, pp. 36–41.

[40] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005, pp. 363–370.

[41] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A Nucleus for a Web of Open Data," in The Semantic Web, ser. Lecture Notes in Computer Science, K. Aberer et al., Eds. Springer Berlin Heidelberg, 2007, vol. 4825, pp. 722–735.

[42] I. Smart TV Alliance, "Technical Specification Version 4.0," 2014, [retrieved: May 2015]. [Online]. Available: https://sdk.smarttv-alliance.org/download.php?file=Smart_TV_Alliance_v4.0_specification.pdf

[43] "HbbTV 2.0 Specification," 2015, [retrieved: May 2015]. [Online]. Available: https://www.hbbtv.org/pages/about_hbbtv/HbbTV_specification_2_0.pdf

[44] R. Cardran, K. Wojogbe, and B. Kralyevich, "The Digital Home: Designing for the Ten-Foot User Interface," 2006, [retrieved: July 2014]. [Online]. Available: http://channel9.msdn.com/Events/MIX/MIX06/BTB029

[45] Samsung, "Design Principles for Creating Samsung Apps Content," 2013, [retrieved: May 2015]. [Online]. Available: http://www.samsungdforum.com/UxGuide/2013/01_design_principles_for_creating_samsung_apps_content.html

[46] D. Loi, "Changing the TV Industry through User Experience Design," in Design, User Experience, and Usability. Theory, Methods, Tools and Practice, ser. Lecture Notes in Computer Science, A. Marcus, Ed. Springer Berlin Heidelberg, 2011, vol. 6769, pp. 465–474.

[47] D. Porta, M. Deru, S. Bergweiler, G. Herzog, and P. Poller, "Building Multimodal Dialogue User Interfaces in the Context of the Internet of Services," in Towards the Internet of Services: The Theseus Program, ser. Cognitive Technologies, W. Wahlster, H.-J. Grallert, S. Wess, H. Friedrich, and T. Widenka, Eds. Springer Berlin Heidelberg, 2014, pp. 149–168.

[48] D. Sonntag, M. Deru, and S. Bergweiler, "Design and implementation of combined mobile and touchscreen-based multimodal web 3.0 interfaces," in Proceedings of the International Conference on Artificial Intelligence, ser. ICAI-09, July 2009, pp. 974–979.

[49] T. Becker, M. Löckelt, C. H. Schulz, S. Bergweiler, M. Deru, and N. Reithinger, "A Unified Approach for Semantic-based Multimodal Interaction," in Towards the Internet of Services: The Theseus Program, ser. Cognitive Technologies, W. Wahlster, H.-J. Grallert, S. Wess, H. Friedrich, and T. Widenka, Eds. Springer Berlin Heidelberg, 2014, pp. 135–148.

[50] S. Bergweiler, M. Deru, and D. Porta, "Integrating a Multitouch Kiosk System with Mobile Devices and Multimodal Interaction," in Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, ser. ITS-2010, ACM. 1515 Broadway New York, New York 10036: ACM, 2010.

[51] W. Wahlster and A. Kobsa, "User Models in Dialog Systems," in User Models in Dialog Systems, ser. Symbolic Computation, A. Kobsa and W. Wahlster, Eds. Springer Berlin Heidelberg, 1989, pp. 4–34.

[52] A. Kobsa, "Generic User Modeling Systems," User Modeling and User-Adapted Interaction, vol. 11, no. 1-2, 2001, pp. 49–63.

[53] N. Reithinger et al., "A look under the hood: design and development of the first SmartWeb system demonstrator," in Proceedings of the 7th international conference on Multimodal interfaces. ACM, 2005, pp. 159–166.

[54] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in Proceedings of the 2008 ACM SIGMOD international conference on Management of data. New York, NY, USA: ACM, 2008, pp. 1247–1250.

[55] O. Hassanzadeh and M. P. Consens, "Linked Movie Data Base," in Proceedings of the Workshop on Linked Data on the Web, LDOW, 2009, [retrieved: May 2015]. [Online]. Available: http://ceur-ws.org/Vol-538/ldow2009_paper12.pdf

[56] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the Accuracy and Robustness of the Leap Motion Controller," Sensors, vol. 13, no. 5, 2013, pp. 6380–6393.

[57] "SPARQL query language for RDF," W3C Recommendation, 2008, [retrieved: May 2015]. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/

[58] "SPARQL 1.1 query language," W3C Recommendation, 2013, [retrieved: May 2015]. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/

[59] S. Bergweiler, "Interactive Service Composition and Query," in Towards the Internet of Services: The Theseus Program, ser. Cognitive Technologies, W. Wahlster, H.-J. Grallert, S. Wess, H. Friedrich, and T. Widenka, Eds. Springer Berlin Heidelberg, 2014, p. 480.

[60] G. Neumann, G. Paaß, and D. van den Akker, "Linguistics to structure unstructured information," in Towards the Internet of Services: The THESEUS Research Program. Springer, 2014, pp. 383–392.

[61] S. Bergweiler, "A Flexible Framework for Adaptive Knowledge Retrieval and Fusion for Kiosk Systems and Mobile Clients," in Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2014), International Academy, Research, and Industry Association (IARIA). IARIA, 8 2014, pp. 164–171.

[62] E. Christensen, F. Curbera, G. Meredith, and S. Weerawarana, "Web Services Description Language (WSDL) 1.1," W3C Note, 2001, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/wsdl

[63] M. Hadley, "Web Application Description Language (WADL)," W3C Member Submission, 2009, [retrieved: May 2015]. [Online]. Available: http://www.w3.org/Submission/wadl

Centricity in Project Risk Management: New Dimensions for Improved Practice

Jose Irizar
School of Business and Management
University of Gloucestershire
Cheltenham, UK
Jose.Irizar@googlemail.com

Martin Wynn
School of Computing and Technology
University of Gloucestershire
Cheltenham, UK
MWynn@glos.ac.uk

***Abstract* – Most organisations engage in major projects during their life cycle, and effective project management is increasingly accepted as a necessary competence in larger companies. Nevertheless, a considerable proportion of projects continue to fail to meet their due dates, exceed budget, do not deliver to specification, miss quality standards, or fall short on customer expectations. The effective management of project risk is a major component of this problem, and central to its resolution; and yet the theory of risk management remains relatively undeveloped and its practice is often poorly executed. This paper examines how the concept of centricity can be applied to some key elements of risk management to develop a conceptual framework that highlights some of the shortcomings of current practice and suggests alternative ways forward. The initial results of applying the model in three major projects in the automotive industry are discussed.**

Keywords - project management; centricity; risk; risk management; risk identification; risk assessment; risk ownership; risk treatment; subjective construct; conceptual model.

I. INTRODUCTION

The quest to improve the management of risk in project implementation has led researchers and practitioners to explore new ways of conceptualizing and classifying risk within project management [1]. Project management is regarded as being of strategic significance in a wide range of industries, and the management of risk is an integral part of the project management process. Despite the recognized criticality of project success for organizations, a considerable proportion of projects continue to either not meet their due dates, exceed budget, do not deliver to specification, fail quality standards, or do not meet customer requirements.

Project failure remains an area of considerable interest in contemporary project management literature, and effective risk management has been identified as one of the major criteria for project success [2]. Yet it remains an area where there is neither a clearly defined theoretical underpinning nor an agreed approach to support the development of a universally agreed method for managing risk. Nevertheless, risk management has become a central component of some of the most widely deployed industry standard methodologies, such as Project Management Body of Knowledge, PRINCE2®, Systems Development Life Cycle, Integrated Capability Maturity Model, and Information Technology Infrastructure Library. Comprehensive risk management is considered as the means by which the effects of unexpected events can be limited, or even how such events can be prevented from happening [3]. Risk management, as an integral component of project management, can thus make a significant contribution to overall project success [4]. This article attempts to develop some new directions in this debate through applying the concept of centricity to a number of themes that run through existing risk management literature – risk identification, risk assessment, risk ownership, and risk treatment. The overall aim of the research is to assess the validity of centricity as a key concept in the development of project management practice. This will also inform policies aimed at enhancing current project risk management, particularly in the automotive industry.

This introductory section is followed by a discussion of the theoretical framework for this paper. The application of the centricity concept to different aspects of risk management is presented in section three, and the base models are further elaborated in section four. Section five then analyses the risk registers of three major projects against these models. Finally, the concluding section summarises results to date and looks at how this research can be further progressed.

II. THEORETICAL FRAMEWORK AND RESEARCH METHOD

The risk management process is often viewed as comprising five main activities [5], and this provides a useful initial frame of reference for this study (Figure 1). Our focus is in the area encompassed by risk identification, risk assessment, risk allocation, and risk control, although our chosen terminology is a little different from that used in this model.

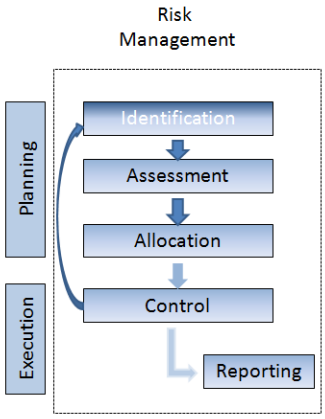


Figure 1. Project Risk Management Process based on PMBoK guide [5]

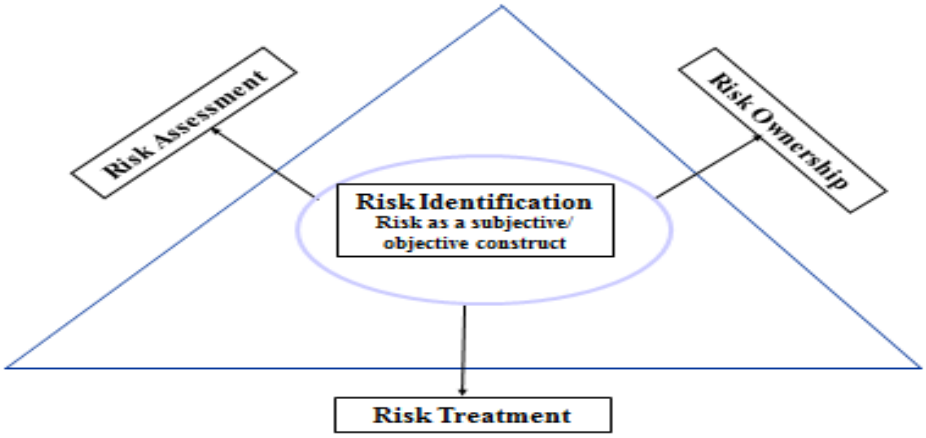


Figure 2. The main dimensions of risk studied using the centricity concept

Risk *identification* is the starting point for risk management in projects. It is considered to be the most influential risk management activity for project outcomes [6], it was found to be one of the most used risk related concepts by organizations [7], and it is recognized by project managers as one of the key areas in need of improvement in complex projects [8]. There are two main schools of thought regarding risk identification – "risk as an objective fact" and "risk as a subjective construct". The former considers risk as epistemologically probabilistic, whilst risk in the subjective construct perspective allows multiple epistemological dimensions of risk [9]. "Risk as an objective fact" considers risks to objectively exist. In the case of "risk as a subjective construct", risk phenomena are subjectively constructed by observers themselves. This subjective-objective construct dichotomy is particularly relevant to the identification of risk, which can also be associated with the concept of "centricity" [1]. Risk as a subjective construct may thus be considered as "person-centric", originating from a subjective perception of risk, rather than from an objective assessment of whether the risk exists and the significance of it.

As regards risk *assessment*, the choice of a particular industry prescribed project management methodology can have a major impact on how risks are assessed, and on overall project outcomes. Project management methodology can be defined as the application of knowledge, skills, tools, and techniques to project activities to meet the project requirements [10] or, using the widest definition given by Cockburn [11], anything that the project management team relies on in order to successfully deliver project results. All of the mainstream methodologies have their own techniques and tools for assessing risks. These methodologies include the Project Management Body of Knowledge (PMBoK), Project Risk Analysis and Management (PRAM), PRINCE2 and the Scrum Agile Standard. The first three of these are generally considered to belong to the so called traditional project management approach, whilst Scrum is the most prominent of the new project management approaches [11].

PMBoK, published by the Project Management Institute (PMI) is the project management guide most widely followed by international organizations. PMI's outreach, its proximity to project management core theories, and its formalization of processes compared to the other standards, make it the optimum standard guide for many authors [12]. One major criticism of PMBoK is its mechanistic approach, making it suitable for routine or technical situations [13], but not so appropriate for unusual or one-off situations. The methodology entails the use of its Probability and Impact Matrix for qualitative risk assessment. Some authors, such as Chapman and Ward [14], challenge the value of this tool for risk assessment. The experience of the risk assessor can determine the so-called probability estimate starting values, and thus estimates become biased. This effect is known as "anchoring" [15].

The development of risk matrices for assessment has taken place isolated from academic research in decision making – risk matrices can produce arbitrary decisions and risk-management actions. These problems are difficult to overcome because they are inherent in the structure of the matrices [16]. Their theoretical basis is superficial and the validity of the qualitative information they employ is highly suspect [17]. The use of risk matrices for assessment illustrates the potential impact of project management methodologies on risk management and project outcomes.

The allocation of *ownership* for identified risks is an essential element of the risk response plan. Ownership is concerned with allocating responsibility for managing project uncertainty to appropriate project parties. Risk

responsibility assignment is considered one of the most influential factors in project risk management success. Risk allocations are fundamental because allocations can strongly influence the motivation of parties and the extent to which project uncertainty is assessed and managed by each party [14]. Recognising that different parties have different objectives, varying perceptions of project risk, and uneven capability for managing associated sources of uncertainty, highlights the significance of risk ownership allocation in the overall risk management process [18].

Looking at the risk ownership allocation process, many risk management professionals see its control as being dependent on the project manager. This leads to the conclusion that the effectiveness of the risk allocation process depends on the project manager's skills, experience and management style. This can be viewed as project-centric risk ownership allocation, with the project manager seen as the key individual in operational delivery of project outcomes. An alternative perspective highlights the possible benefit of assigning risk ownership allocation control, and risk allocation itself, to a range of individuals, who may not be in regular contact with the project manager [14]. Practitioners' responses suggest that an alternative system that encourages all project members to participate in the risk management process is normally missing. A consequence may be the failure to create a collective responsibility to manage risk [18].

A further risk dimension discussed here is risk *treatment*. Project risk treatment is the stage at which the risk strategy is defined. The strategy defines how to manage risk. This can be anywhere in a spectrum from reduce exposure or mitigate impact to transfer/externalize risk or accept risk. The decision to choose any of these responses can be supported by tools that provide risk factor dependencies and priorities [19]. Risk treatment thus depends on the risk propensity or attitude to taking risks. Behaviour towards taking risks may change over time through education, training and experience. A balanced risk treatment will probably increase the threshold at which point the organization is willing to take risks. As a result, an organization may enhance its competitive edge. If it is averse centric in its treatment of risk, it will be less likely to take risks, having a lower propensity for risk taking.

A balanced approach to risk treatment would be one focusing both on risk and reward. An overemphasized focus on risk versus reward may have considerable influence on strategic decisions such as entering new markets, developing new products or targeting new mergers and acquisitions [20].

Executive inaction may result in loss of potential revenue growth. Education and training in project risk management, with subsequent additional experience in the organization, may lead to a better understanding of risk and reward. People themselves are a major source of risk, and education, training and experience can make them part of the solution. Proper risk management can be seen as a protective shield for the organisation, rather than an action stopper. Management and employees together learn through education and training to take and manage risks, not to avoid them. The organization can thereby treat risk appropriately and not circumvent it.

The aim of this research is to explore how the concept of centricity can be applied to the four dimensions of risk management discussed above (see Figure 2). Centricity in a managerial context can be defined as the mind set or attitude that characterises the managers' or organisation's outlook and motivation in the relationship to others [21] [22]. In recent years, qualitative research has found increasing recognition in many areas of project management practice. A large number of empirical studies using qualitative data are available in academic literature and specialized journals [2] [4] [9]. At the same time, management researchers and practitioners in particular rely on evidence-based policy. In fact, most of the existing generally accepted standards in the project management field as a whole are built around evidence-based policy and best practice.

Through an analysis of existing literature, allied to empirical data and observations in large project environments, this paper looks to develop a conceptual framework for research in the following areas:

- Person-centric risk identification vs. objective risk identification
- Methodology-centric risk assessment vs. multi-disciplinary/eclectic risk assessment
- Project-centric risk ownership allocation vs. devolved ownership allocation
- Averse-centric risk treatment vs. balanced risk treatment

This approach assumes that it is feasible and sensible to cumulate findings and generalize results to create new knowledge. The application of the centricity concept to the aspects of risk management discussed in this paper will be tested and developed further through primary research case studies as part of an on-going research project.

III. CONCEPT DEVELOPMENT

The identification of risk as a subjective phenomenon coincides with its creation – the risk exists only once the stakeholder has identified it. This is particularly noticeable for risks linked to an organization's own qualities and deficiencies [23]. This subjective or person-centric risk identification can often produce inefficiencies in the management of risk that may impact detrimentally on project cost and overall project success (see Figure 3).

The analysis of risks associated with different information systems (IS) by Ward and Griffiths [24] uses a strategic grid depiction of risk categories (Figure 4) that can be used in the application of the centricity concept for project risk management. If we view risk identification against risk assessment in grid format, many projects - arguably the majority - adopt a person-centric approach to risk identification and a methodology centric approach to risk assessment. Yet we suggest, as an initial standpoint, that a combination of objective risk identification and eclectic risk assessment is likely to produce the most successful project outcomes (see Figure 5).

The use of risk matrices for risk assessment illustrates this well. Their apparent simplicity and transparency are reasons for their popularity; however, they potentially entail

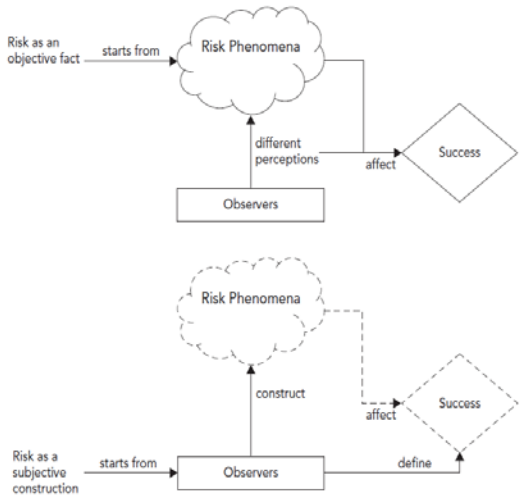


Figure 3. The two means of risk identification [9].

serious mathematical defects and inconsistencies. Different risk assessors may assign greatly different ratings to the same risk exposure [25]. Such different ratings are due to fundamentally different worldviews, beliefs, and other psychosocial factors, the consequences of which are not significantly changed through reflection and learning.

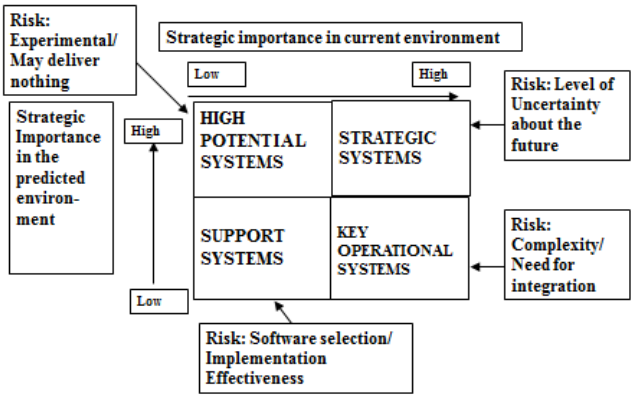


Figure 4. Quadrant grid depiction of IS risk categories [24]

There are a number of evident shortcomings in the use of these matrices. These include instability resulting from categorization differences, and the lie factor, which suggest that they can obscure rather than enlighten communication. The rankings produced have been shown to be unduly influenced by the matrix design, which is ultimately arbitrary. It is suggested that other means of assessing risk based on decision-analytical methods could produce improved outcomes [17]. An example of a decision-making

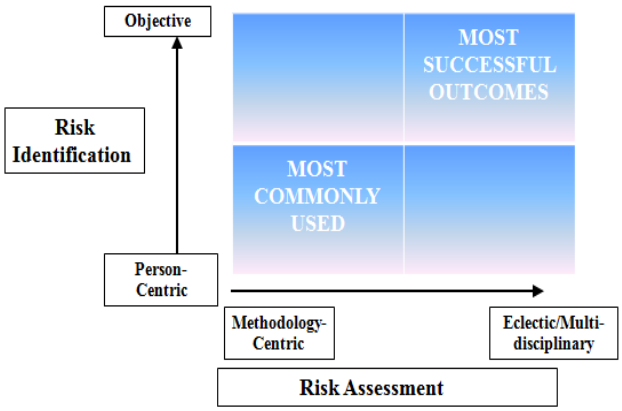


Figure 5. Risk identification and risk assessment: basic model

tool applicable to new product development (NPD), designed to help the project manager choose the best way to improve project success rates while controlling the level of risks, is presented by Marmier, Gourc and Laarz [26]. Other authors combine content analysis with cluster analysis of existing historical data, to develop the Risk Breakdown Structure which can be used to build risk management guidelines [27]. These scientific decision analysis tools could be an alternative to the popular but inefficient use of risk matrices for risk prioritization. The establishment of systematically maintained lessons learned datasets could also provide quantitative reliable data to estimate the likelihood of potential events.

There are some similarities in an initial assessment of risk identification and risk ownership using the centricity concept (see Figure 6). Risk ownership centricity is viewed as an overdependence on centralised control and allocation of risks, and their subsequent management and resolution. The different approaches to the ownership of risk management often appear as a conflict between centralized project risk management and the empowerment of sub-project teams [28]. The complexity of certain projects makes it difficult to understand the consequences of central decisions for the team members. The project manager alone will struggle to comprehend the details of all potential risks, oversee these and control their management. Yet many projects are project centric in terms of risk management process and person-centric as regards risk identification. The on-going monitoring and maintenance of the risk register in which project risks are listed is often controlled by the central project manager [29]. It is suggested that overall project outcomes would be improved by appropriately combining centralized and decentralized ownership of risk management, especially in complex projects. More particularly, project management practitioners in industries which require intense collaboration - such as automotive product development - complain about the insufficient development of risk management methods and processes not being integrated and synchronized. Lack of

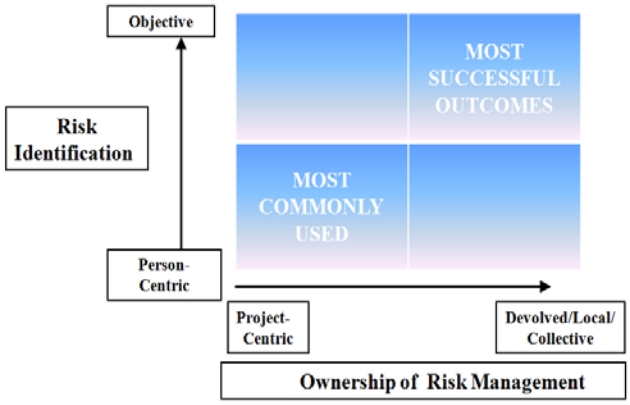


Figure 6. Risk identification and ownership: basic model

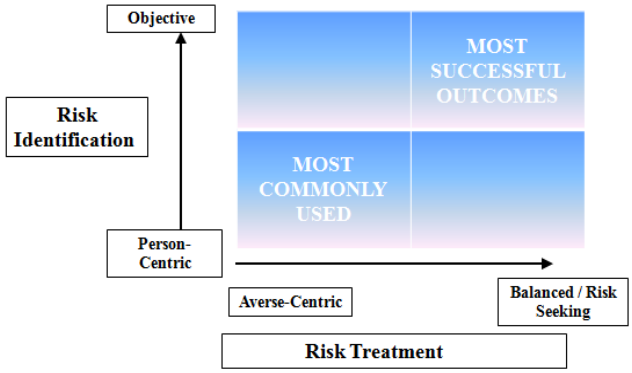


Figure 7. Risk identification and risk treatment: basic model

collaborative risk management, together with poor communication, is the main reason for project failure in the automotive industry [30].

Similarly, in major IS projects, the IT function has traditionally owned and led information risk management and security operations. However, the move to user ownership of systems requirements, process improvement issues and data access and maintenance, have changed the risk and security paradigm. Business managers, systems users and the IT function are now required to understand and learn others risk-reward trade-offs. The IT function must now share ownership of the risk management process and transfer accountability for some key areas of risk to business partners [31].

The final dimension considered here is risk treatment, again juxta positioned against the central theme of risk identification (Figure 7). As noted above, centricity in a managerial context can be viewed as a mind-set or attitude that characterises the managers' or organisation's outlook and motivation in their relationship with others. Averse centric organizations will be less likely to take risks in their treatment of risk as they show a lower propensity for risk taking. Risk averse organisations may even avoid managing risks or limit resources available for risk management activities, which will work against effective risk management making these organisations, paradoxically, more vulnerable to risk [32].

IV. Model Progression and Implications

The basic conceptual model can be developed further in the light of literature analysis and project experience, indicating the downsides and upsides of operating in each quadrant of the model (see Figures 8, 9 and 10). This may also have implications for the use of some of the mainstream project management methodologies in their treatment of risk issues.

For example, PMI's project management guide, although considered as the best in class among all available methodologies and guides, could be enhanced with some early risk identification tools and techniques from more minor project management methodologies such as Scrum.

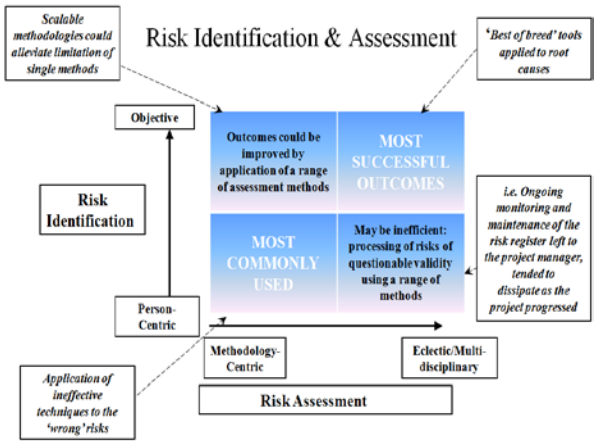


Figure 8. Risk identification and risk assessment: model development

Such enhancements would help reduce project uncertainty. In addition, experience gained by specific industries' customized methodologies can increase risk management effectiveness. These could provide quantitative data to support estimations of the probabilities of risks occurring. Equally, decision analysis tools are an alternative or complement to the inconsistent but widely used risk matrices. Decision analysis tools may be initially difficult to adopt; however, they can provide objective data to support risk assessment as an alternative to the use of risk matrices with all their inherent deficiencies.

The popularity of new project management approaches, such as that embodied in Scrum, resides in their adaptability to accommodate change and the unexpected, as opposed to the quest for risk predictability, which is the basis of the traditional approaches [12]. These new approaches also highlight the importance of both formal and informal communication, collaboration between project team members, and their involvement in decision making, suggesting that a more devolved and collective risk management process is generally beneficial.

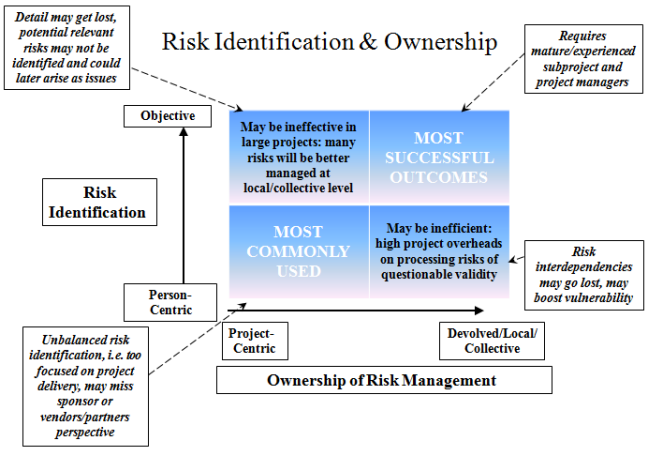


Figure 9. Risk identification and risk ownership: model development

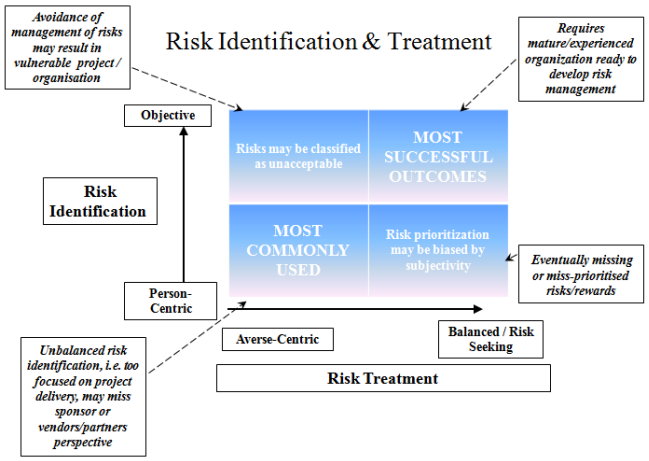


Figure 10. Risk identification and risk treatment: model development

Context, such as the project organization's size and complexity, may play a significant role in tailoring and adapting any project risk management methodology when applying the different standards. Generally speaking, the traditional approach is more appropriate for projects with a very low level of uncertainty in which emphasis will be on planning. Conversely, agile project management, with a more flexible approach to a collective risk management process, fits best in environments characterized by a high level of uncertainty [12].

The two standards with a greater emphasis upon early risk identification are PRINCE2 and Scrum. Traditional project management practices struggle to deal effectively with uncertainty. In highly uncertain environments, approaches such as Scrum and lean methods can help manage residual uncertainty about risk not addressed by traditional project management practices [33].

The model developed using centricity concepts suggests that a combination of risk management based on traditional standards and more flexible approaches typified by Scrum would be beneficial for most projects. However, this would imply significant mindset changes in the organisation [34]. Project teams need to be empowered to effectively use a range of different methodologies and techniques, which may involve team members adopting new roles. This may result in teams creating their own, tailored, risk management process and activities [35].

V. PRELIMINARY RESEARCH RESULTS

The provisional conceptual framework has been used in assessing risk in three major projects in the automobile sector. Two of these projects relate to the implementation of a mainstream ERP packaged software product (projects 1 and 2); and the third (project 3) concerns the development of a mechanical steering gear product for an international automotive Original Equipment Manufacturer (OEM). The main source of data has been the risk registers in these three projects, which detail 15, 20, and 48 risks respectively.

These risks were first classified against risk identification and risk assessment criteria (Table I and Figure 11). The majority of risks fall into the top left quadrant (Quadrant 1), signifying objective identification of risk, and methodology centric in terms of assessment. Most risks can be considered to have been objectively identified - after discussion and agreement with the project manager and other colleagues. Yet there remains a degree of subjectivity in most risk identification, particularly in the first ERP project, where many risks were registered individually by the team member or group. Although these were validated or completed by the project manager, there was still a certain degree of subjectivity in the item description, its cause, assumptions, probability estimation or estimated impact on objectives.

This can be better understood by looking at the five risks from Project 3 that fall in Quadrant 4 in Figure 11 - where risk identification is adjudged person centric and risk assessment remains methodology centric. Four out of these five risks can be classified as "project schedule risks" (where timescale is a major uncertainty), and the fifth one can be classified as a "specification risk", (where completeness of specification is at risk). A lack of collective, objective assessment is indicated by the fact that, in the risk register, the risk type or risk category was not adequately maintained or updated by the project manager or any other team member during the project life cycle; and once the countermeasures agreed to mitigate the risk items were completed, these risks were then eliminated from the register without adequate consideration. From the risk register, examples of "project schedule risks" include "risk of delay in design verification due to component prototype timing" and "potential misalignment between supplier key product characteristics matrices". In the first example, once the manufacturing team had confirmed the prototype timing was not an issue for design verification, the risk item was closed. In a similar manner, for the second item, after the engineering representatives confirmed that there was no misalignment between the two lists with the responsible suppliers, the risk item was closed, the result of this confirmation being risk "elimination". These are examples of how person centric risk identification and methodologically centric risk assessment

can combine to produce decisions that may be neither properly objective nor likely to engender sound project management outcomes.

Risks were then mapped against risk identification and risk ownership. More mature organizations may deal with risk in a more devolved manner – sub-project teams may be accustomed to having exposure to risks and have the knowledge and experience to manage them effectively. There was some evidence of this in the project to develop the mechanical steering gear product (project 3 – see Table II). The project team members and the project stream leads or sub-project leads were experienced enough to identify, record and suggest counter measures to a small number of risks, which were managed in this way (the 2% in Quadrant 2 in Figures 12 and 13).

These two risk items were managed by the engineering sub-group with no or minor involvement from the project manager. They represent two objectively identified risks that were owned and managed in a devolved/local manner. The risks were associated with two new components which failed two critical quality criteria - process validation and design verification. Both risk items reflected a lack of experience in the organization in general regarding the design and conception of the mentioned components. The engineering sub-group arguably had most experience in managing projects and dealing with risks. Counter measures suggested and pursued for the management of these risks were early sourcing, early involvement of suppliers in the design process, and adequate testing using an accepted standard – the Failure Mode and Effect Analysis (FMEA) process. The majority of risks across all three projects were, however, largely owned by the central project team. Organizations with less of a project management culture are more dependent on the project manager skills when dealing with risks.

A classification of risks on the risk identification-risk treatment axes indicates that a balanced attitude to risk taking was prevalent across all three projects. This reflects the relatively mature nature of these organisations, where calculated risk-taking is recognized as an element of overall management. The fact that the vast majority of the risks were identified *after* project approval in itself indicates a confidence in these organisations that all groups involved are able to work together to develop a response plan to deal with identified risks. This is reflected in Figure 14, indicating that all risks, however identified, were dealt with in a "balanced" or "risk seeking" manner, as opposed to risk averse centric.

VI. CONCLUDING REMARKS

This article has explored how the concept of centricity can be applied to some key aspects of project risk management to aid understanding and develop alternative perspectives. The concept of centricity has been used as a key component in the development of a conceptual model that is now being tested and refined through primary case study research of risk management processes in IS and new product development projects in the automotive industry. This entails action research, through which the conceptual framework is being applied and developed in major pan-

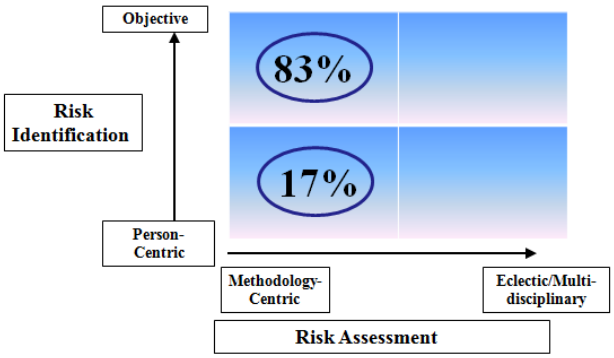


Figure 11. Risk identification and risk assessment: preliminary mapping of first research results
(Quadrant 1 – top left; Quadrant 2 – top right; Quadrant 3 – bottom right; Quadrant 4 – bottom left)

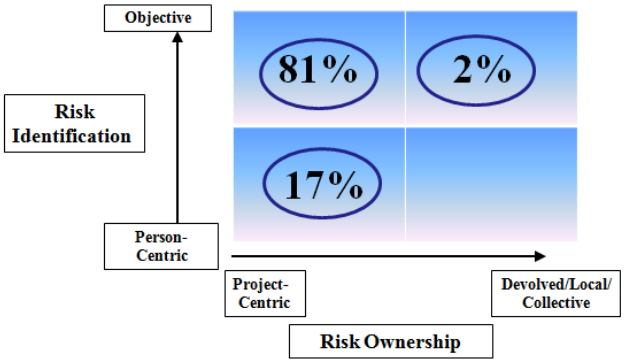


Figure 12. Risk identification and risk ownership: preliminary mapping of first research results

European projects.

Our initial assumptions were that in most projects, risk identification is person-centric, risk assessment is methodology-centric, and the overall risk management process is project-centric. Yet current literature, recent trends and personal observation suggest that a move away from centricity in these components of risk management would benefit project outcomes. The integration of traditional and agile project management methodologies and their tailoring to the specific needs of the organization is gaining wide practitioner and academic attention. The initial results from primary research case studies in organisations with a strong management culture and significant project experience generally support the initial assumptions. However, they also raise a number of issues that are now being pursued through more detailed analysis of each of the three cases. The various dimensions of risk management will be matched against different aspects of each project – project focus, duration, budget, resourcing, ownership, expectation, and tolerances for example – as well as with project outcomes; and a wider range of more in-depth interviews is being conducted to

TABLE I. RISK IDENTIFICATION AND ASSESSMENT IN THE THREE PROJECTS: QUADRANT ALLOCATION

Risk identification and assessment				
Project 1	**Quadrant1**	**Quadrant 2**	**Quadrant 3**	**Quadrant 4**
Total	13			7
Total %	65%			35%
Project 2	1	**2**	**3**	**4**
Total	13			2
Total %	87%			13%
Project 3	1	**2**	**3**	**4**
Total	43			5
Total %	90%			10%
TOTAL	1	**2**	**3**	**4**
Total	69			14
Total %	83%			17%

widen perspectives and more firmly ground assessments in first hand interview material. Once this further research stage is completed, the contribution of the centricity concept to improved risk management practice will be clearer, but initial research results suggest that this is a new way of looking at risk management that can add value to the overall process.

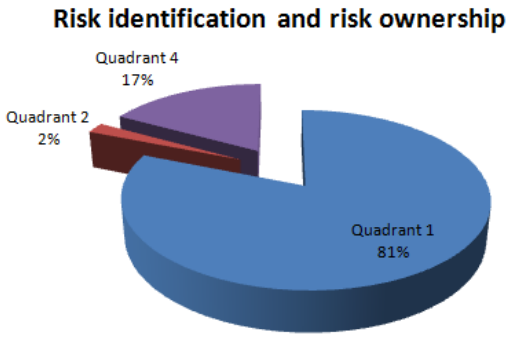


Figure 13 . Risk identification and risk ownership: quadrant allocation

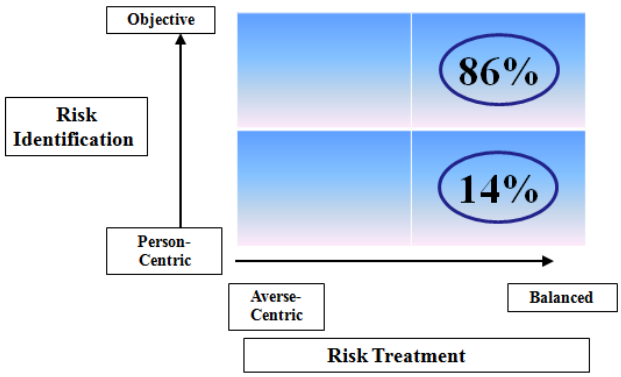


Figure 14. Risk identification and risk treatment: preliminary mapping of first research results

TABLE II. RISK IDENTIFICATION AND OWNERSHIP IN THE THREE PROJECTS: QUADRANT ALLOCATION

Risk identification and ownership				
Project 1	**Quadrant1**	**Quadrant 2**	**Quadrant 3**	**Quadrant 4**
Total	13			7
Total %	65%			35%
Project 2	**1**	**2**	**3**	**4**
Total	13			2
Total %	87%			13%
Project 3	**1**	**2**	**3**	**4**
Total	41	2		5
Total %	85%	4%		10%
TOTAL	**1**	**2**	**3**	**4**
Total	67	2		14
Total %	81%	2%		17%

This is illustrated by the challenge facing the project manager considering how to manage overall risk. The question is not just which project management risk approach should be adopted, but more how to select a “best of breed approach”, choosing the most suitable techniques, templates,

tools and artifacts out of the different standards and methodologies that are available. It is hoped that this research, by introducing the concept of centricity to analyse current practice, will engender this process and lead to better overall project outcomes. As Peter Drucker has put it, "when intelligent, moral, and rational people make decisions that appear inexplicable, it's because they see a reality different to the one seen by others" [36]. This phenomenon, in the case of risk management, requires further research into the interaction and communication between individuals, project teams and their contexts. If the centricity concept can be successfully harnessed to underpin this research, it has the potential to significantly enhance eventual project outcomes.

REFERENCES

[1] J. Irizar and M. Wynn, "Centricity in Project Risk Management: Towards a Conceptual Framework for Improved Practice," CENTRIC 2014: The Seventh International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, Nice, France, October 12-16, 2014, pp. 83-88, ISBN: 978-1-61208-369-8.

[2] D. McClure, From the CIO trenches: Why some projects fail and others succeed, Gartner Industry Research, 2007.

[3] R. Jen, Visual Ishikawa Risk Techique (VIRT) - An approach to risk management, PMI Virtual Library, 2009, URL: http://www.pmi.org/en/Knowledge-Center/Knowledge-Shelf/~/media/Members/Knowledge%20Shelf/Jen_2009.ashx [accessed: 2015-05-20].

[4] K. de Bakker, "Risk management affecting IS/IT project success through communicative action," Project Management Journal, 42(3), 2011, pp. 75-90.

[5] V. Holzmann, Analyzing Lessons Learned to Identify Potential Risks in new Product Development Projects, Paper presented at the 6th European Conference on Information Management and Evaluation, 2012, pp. 127-134.

[6] K. de Bakker, A. Boonstra, and H. Wortmann, "Risk managements' communicative effects influencing IT project success," International Journal of Project Management, 30(4), 2012, pp. 444-457.

[7] P. L. Bannerman, "A Reassessment of Risk Management in Software Projects" in Handbook on Project Management and Scheduling, Vol. 2, C. Schwindt & J. Zimmermann, Eds., Switzerland, Springer International Publishing, 2015, pp. 1119-1134.

[8] C. M. Harvett, "A Study of Uncertainty and Risk Management Practice Related to Perceived Project Complexity," (PhD), 2013, URL: http://epublications.bond.edu.au/theses/73/ [accessed: 2015-05-20].

[9] H. Zhang, "Two schools of risk analysis: A review of past research on project risk," Project Management Journal, 42(4), 2011, pp. 5-18.

[10] PMI, A guide to the project management body of knowledge (PMBOK®) (Fifth ed.) Project management institute, Inc., 2013.

[11] M. Špundak, "Mixed Agile/Traditional Project Management Methodology – Reality or Illusion?" Procedia - Social and Behavioral Sciences, 119, 2014, pp. 939-948.

[12] M. J. Thaheem, "Project Risk Management for Sustainable Restoration of Immovable Cultural Heritage: Lessons from Construction Industry and Formulation of a Customized PRM Model," (Doctorate of Philosophy), 2014, URL: http://porto.polito.it/2531894/1/THAHEEM_Tesi.pdf [accessed: 2015-05-20].

[13] P. W. G. Morris, L. Crawford, D. Hodgson, M. M. Shepherd, and J. Thomas, "Exploring the role of formal bodies of knowledge in defining a profession – The case of project management," International Journal of Project Management, 24(8), 2006, pp. 710-721.

[14] C. Chapman and S. Ward, Project risk management: processes, techniques and insights, John Wiley & Sons, 2003.

[15] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," Science, 185(4157), 1974, pp. 1124-1131.

[16] P. Thomas, "The Risk of Using Risk Matrices," Masters thesis, University of Stavanger, 2013, URL: http://brage.bibsys.no/uis/bitstream/URN:NBN:no-bibsys_brage_45899/1/Thomas_Philip.pdf [accessed: 2015-05-20].

[17] K. D. Wall, "The Trouble With Risk Matrices," DRMI Working Papers Ongoing Research, 2011, pp. 11-23.

[18] C. M. Harvett, "A Study of Uncertainty and Risk Management Practice Related to Perceived Project Complexity," PhD thesis, 2013, Bond University, ePublications@bond.

[19] D. Aloini, R. Dulmin, and V. Mininno. "Risk assessment in ERP projects," Information Systems, 37(3), 2012, pp. 183-199, doi: 10.1016/j.is.2011.10.001.

[20] TowerGroup, "Reducing Risk Management's Organizational Drag Executive Guidance," Vol. Executive Guidance Q3 2014, Arlington VA, CIO Executive Board, 2014.

[21] H. V. Perlmutter, "The Tortuous Evolution of the Multinational Corporation," Columbia Journal of World Business, 4 (1), 1969, pp. 9-18.

[22] M. Olsen and A. Roper, "Towards an Understanding of Centricity: Profiling International Hotel Groups," International Journal of Hospitality Management, 17(2), 1998, pp. 111-124.

[23] J. Irizar and M. Wynn, "Risk as a Subjective Construct: Implications for Project Management Practice," The Fifth International Conference on Information, Process, and Knowledge Management (eKNOW 2013) IARIA, Feb. 2013, pp. 135-141, ISBN: 978-1-61208-254-7.

[24] J. Ward and P. Griffiths, Strategic planning for information systems, 2nd ed., Chichester, John Wiley& Sons, 1996.

[25] D. J. Ball and J. Watt, "Further Thoughts on the Utility of Risk Matrices," Risk Analysis, *33* (11), 2013, pp. 2068-2078. doi: 10.1111/risa.12057.

[26] F. Marmier, D. Gourc, and F. Laarz, "A risk oriented model to assess strategic decisions in new product development projects," Decision Support Systems, 56, 2013, pp. 74-82.

[27] V. Holzmann, Analyzing Lessons Learned to Identify Potential Risks in new Product Development Projects, Paper presented at the 6th European Conference on Information Management and Evaluation, 2012, pp. 127-134.

[28] T. M. Williams, "Empowerment vs risk management?" International Journal of Project Management, 15, 1997, pp. 219-222.

[29] P. L. Bannerman, "Risk and risk management in software projects: A reassessment," The journal of systems and software, 81(12), 2008, pp. 2118-2133.

[30] K. Niebecker, "Collaborative and cross-company project management within the automotive industry using the Balanced Scorecard," Ph.D. dissertation, Faculty of Engineering and IT, University of Technology, Sydney, 2009.

[31] E. Chobanova, "Why You Should Share Your Risk With Business Partners," [Online], URL: http://www.executiveboard.com/it-blog/why-you-should-share-your-risk-with-business-partners, 2014, [accessed: 2015-05-20].

[32] P. L. Bannerman, “A Reassessment of Risk Management in Software Projects,” Handbook on Project Management and Scheduling Vol. 2, C. Schwindt & J. Zimmermann, Eds. Springer International Publishing, 2015, pp. 1119-1134.

[33] C. Besner and B. Hobbs, “The paradox of risk management; a project management practice perspective,” International journal of managing projects in business, 5 (2), 2012, pp. 230-247.

[34] M. McWha, 2014, “Agile is a Mindset, not a Methodology,” URL: http://www.executiveboard.com/it-blog/agile-is-a-mindset-not-a-methodology [accessed: 2015-05-20].

[35] R. Rodríguez Gutiérrez, J. Minguella Canela, F. Fenollosa i Artés, B. Ventayol Femenias, and M. A. d. l. Santos López, “Experiences in Agile R&D Project Management for New Product Design and Development in the Automotive Industry,” The 16th International Research/Expert Conference on Trends in the Development of Machinery and Associated Technology, TMT 2012, pp. 223-226.

[36] Bud Baker, “The fall of the firefly: An assessment of a failed project strategy,” Project Management Journal, 33 (3), 2002, pp. 53-57.

PRINCE2® is a Trade Mark of The Office of Government Commerce.

Process Mining in the Education Domain

Awatef HICHEUR CAIRNS[1], Billel GUENI[1], Mehdi FHIMA[1], Andrew CAIRNS[1] and Stéphane DAVID[1]
Nasser KHELIFA[2]

[1] ALTRAN Research, [2] ALTRAN Institute
2 rue Paul Dautier
Vélizy-Villacoublay, 78140 FRANCE
[awatef.hichaurcairns, billel.gueni, mehdi.fhima, andrew.cairns, stephane.david, nesser.khelifa]@altran.com

***Abstract*— Given the ever changing needs of the job markets, education and training centers are increasingly held accountable for student success. Therefore, education and training centers have to focus on ways to streamline their offers and educational processes in order to achieve the highest level of quality in curriculum contents and managerial decisions. Educational process mining is an emerging field in the educational data mining (EDM) discipline, concerned with developing methods to better understand students' learning habits and the factors influencing their performance. It aims, particularly, at discovering, analyzing, and providing a visual representation of complete educational processes. In this paper, in continuity of the work presented in [1], we investigate further the potential, challenges and feasibility of the educational process mining in the field of professional trainings. First, we focus on the mining and the analysis of social networks, from educational event logs, between courses units, resources or training providers. Second, we propose a clustering approach to decompose educational processes following key performance indicators. We have experimented this approach using the ProM Framework.**

Keywords-component; process mining; educational data mining; curriculum mining; key performance indicator; ProM.

I. INTRODUCTION

Recently, education and training centers have started introducing more agility into their teaching curriculum in order to meet the fast-changing needs of the job market and meet the time-to-skill requirements. Modern curriculums are no longer monolithic processes. Students can pick the courses from different specialties, may choose the order, the skills they want to develop, the level (from beginner to specialist), the way they want to learn (theoretical or practical aspects) and the time they want to spend. This need for personalized curriculum has increased with the emergence of collaborative tools and on-line training which often supplement and sometimes replace traditional face-to-face courses [2]. In fact, e-learning represents an increasing proportion of the in-company training, while addressing ever wider populations. The broad number of courses available and the flexibility allowed in curriculum paths could create, as a side effect, confusion and misguidance. Students may be overwhelmed by the offer and blurred on the time required to enter and remain in the job market. Moreover, teachers and educators may lose control of the education process, its end-results and feed-back [2]. In this modern education context, where students can access courses and curriculums on-line, from all around the world, the education systems enter a competitive market they are not used to, where they are increasingly held accountable for students' success. This situation creates additional pressure in higher educational institutions and training centers to achieve the highest level of quality in curriculum content and managerial decisions.

The use of information and communication technologies in the educational domain generates large amount of data, which may contain insightful information about students' profiles, the processes they went through and their examination grades. The deriving data can be explored and exploited by the stakeholders (teachers, instructors, etc.) to understand students' learning habits, the factors influencing their performance and the skills they acquired [3-4]. Rather than relying on periodic performance tests and satisfaction surveys, exploiting historical educational data with appropriate mining techniques enables in-depth analysis of students' behaviors and motivations. To answer these questions, there are increasing research interests in using data mining in education [3-4]. *Educational Data Mining* (EDM) is a discipline aimed at developing specific methods to explore educational datasets generated by any type of information system supporting learning or education (in schools, colleges, universities, or professional training institutions providing traditional and/or modern methods of teaching, as well as informal learning). EDM brings together researchers and practitioners from computer science, education, psychology, psychometrics, and statistics. EDM methods can be classified into two categories – (1) Statistics and visualization (e.g., Distillation of data for human judgment), and (2) Web mining (e.g., Clustering, Classification, Outliers detection, Association rule mining, Sequential pattern mining and Text mining) [5]. However, most of the traditional data mining techniques focus on data or simple sequential structures rather than on full-fledged process models with concurrency patterns [6-7]. For instance, it is not clear how, given a study curriculum, EDM techniques could check automatically whether the students always follow it [6]. Precisely, the basic idea of *process mining* [8] is to discover, monitor and improve real processes (i.e., not assumed nor truncated processes) by extracting knowledge from event logs recorded automatically by Information Systems. Our research aims to develop generic methods which could be applied to general education issues and more specific ones concerning professional training or e-learning fields for [1], [9]:

- The extraction of process-related knowledge from large education event logs, such as: process models

and social networks following key performance indicators or a set of curriculum pattern templates.
- The analysis of educational processes and their conformance with established curriculum constraints, educators' hypothesis and prerequisites.
- The enhancement of educational process models with performance indicators: execution time, bottlenecks, decision point, etc.
- The personalization of educational processes via the recommendation of the best course units or learning paths to students (depending on their profiles, their preferences or their target skills) and the on-line detection of prerequisites' violations.

In this paper, we focus mainly on (1) process model discovery, deriving from Key Performance Indicators; (2) social network discovery between training courses and training providers. For the first time, to our knowledge, the present approach handles a professional training dataset of a consulting company involved in the training of professionals. In this paper, we extend the work presented in [1]. The rest of this paper is organized as follows. Section II introduces process mining techniques and their application in the educational domain. Section III presents our approach for social networks mining and process models discovery. Section IV describes briefly the PHIDIAS platform. Section V discusses some related works. Finally, Section VI concludes the paper.

II. EDUCATIONAL PROCESS MINING

A. Definition

Process mining is a relatively new technology which emerged from business community [8]. It focuses on the development of a set of intelligent tools and techniques aimed at extracting process-related knowledge from event logs. The complete overview of process mining application in the educational field (known as *educational process mining* [6-7]) is illustrated in Fig. 1.

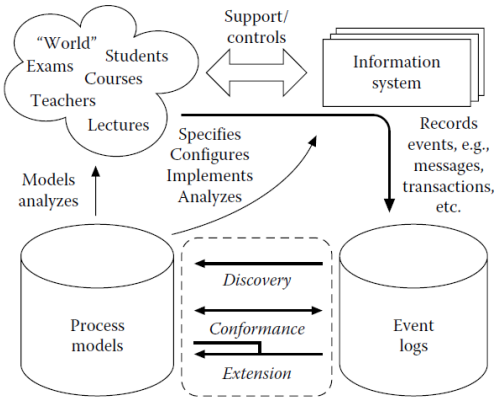


Figure 1. Process mining concepts

An event log corresponds to a set of process instances (i.e., traces) following a business process. Each recorded event refers to an activity and is related to a particular process instance. An event can have a timestamp and a performer (i.e., a person or a device executing or initiating an activity). Typical examples of event logs in education may include students' registration procedures and attended courses, student's examination traces, use of pedagogical resources and activity logs in e-learning environments. The three major types of process mining techniques are:

Process model discovery: takes an event log and produces a complete process model able to reproduce the behaviour observed in this log. Examples of such techniques are control-flow mining algorithms, which allow the discovery of educational processes and learning paths based on the dependency relations that can be inferred from event logs, among student's actions or courses taken. This step is not limited to control-flow discovery. For instance, there are approaches to discover social networks, organizational structures and resource behavior from event logs. Typically, these approaches use the information about the performer (resource), e.g., the person or component initiating or completing some activity, to generate the relationships in social networks between these resources, following their involvement in the process execution [10].

Conformance checking: aims at monitoring deviations between observed behaviours in event logs and normative process models (generated either by traditional modelling or by process mining techniques). This evaluation can be made using metrics such as *fitness* (Is the observed behavior possible according to the model?) and a*ppropriateness* (Is the model *typical* for the observed behavior?). Major deviations from a normative model might also mean that the model itself does not reflect real world circumstances and requirements. *Compliance checking* is another kind of event log inspection, which aims at measuring the adherence of event logs with predefined business rules, constraints, temporal formulas or Quality of Service (QoS) definitions. An example of such a technique for auditing event logs is the LTL Checker which is used to verify properties (i.e., rules, constraints, etc.), expressed in terms of Linear Temporal Logic (LTL).

Process model extension: aims to improve a given process model based on information (e.g., time, performance, case attributes, decision rules, etc.) extracted from an event log related to the same process. There are different ways to extend a given process model with these additional perspectives, e.g., decision mining, performance analysis, and user profiling.

Process mining techniques can also be used for operational decision support activities. For instance, based on historic information, it is possible to make predictions (e.g., the remaining flow time) for running cases or to recommend suitable actions (e.g., proposing the activity that

will minimize the expected costs and time). Moreover, it is possible to check, on the fly, if running cases fit with normative process models or if desired properties (defined in Linear Temporal Logic) hold in these running cases.

B. The PROM Framewok

Regarding available process mining tools, the ProM Framework [11] is the most complete and powerful one aimed at process analysis and discovery from all perspectives (process, organizational and case perspective). It is implemented as an open-source Java application with an extendable pluggable architecture, which enables users to write and import their own mining algorithms as plug-ins. These plug-ins can also be chained into `macro' plugins. ProM supports a wide range of techniques for process discovery, conformance analysis and model extension, as well as many other tools like conversion, import and export plug-ins. The de facto standard for storing and exchanging events logs is the MXML (Mining eXtensible Markup Language) format or more recently the XES (eXtensible Event Stream) format, which is the successor of MXML. These two standards are adopted in the ProM Framework. In practice, however, ProM presents certain issues of flexibility and scalability, which limit its effectiveness in handling large logs from complex industrial applications [12]. We may get over these limitations by using the service oriented architecture of the ProM 6 framework. Theoretically, such architecture may allow the distribution of ProM's plugins over multiple computers (e.g., grid computing). We are recently testing such a construction in the development of an interactive and distributed platform tailored for educational process discovery and analysis.

C. Process Mining Issues in the Education Domain

Educational systems support a large volume of data, coming from multiple sources and stored in various formats and at different granularity levels [3-4]. The data come from face to face educational systems, such as traditional classrooms, or from distance education taken from interactive learning environments or computer-supported collaborative learning. For instance, face to face educational systems store only administrative and demographic information; i.e., students' profiles (e.g., grades and curriculum goals), who follows which program, takes which courses and exams. Computer and on-line education systems store more fine-grained data because they can record all the information about students' actions and interactions into log files and databases. This data includes resource usage logs (e.g., handouts, video recordings), assessment data, collaborative writing in wikis or versioning systems, and participation in forums [2]. Moreover, the cost-effectiveness quest of modern education systems leads them to record more information about (1) the learners' short-term satisfaction on programs, course units or resources, and (2) the long-term usefulness of the courses they have taken in entering and remaining in employment. Recorded information in educational systems are structured (logs, student registration information, student usage profiles, administrative information, etc.) or unstructured (interaction with teachers via chat, collaboration with other students via chat, etc.).

To discover a suitable process model, it is assumed that the event log contains representative sample of behavior. However, the application of process discovery techniques presents some challenges given the huge volume and the traces' heterogeneity often encountered in educational datasets.

Voluminous Data - Large Number of Cases or Events in event logs

Event logs in the education domain, particularly those coming from e-learning environments, may contain massive amounts of fine granular events and process related data. In fact, real-life experiences show that most of the contemporary process mining techniques/tools are unable to handle massive event logs [12], [14]. There is a need for research in both the algorithmic as well as deployment aspects of process mining. For example, we should move towards developing efficient, scalable, and distributed algorithms in process mining [12]. This issue was tackled in recent researches [14], [15], where clustering techniques were proposed for partitioning large logs into smaller parts that can be checked locally and more easily.

Heterogeneity and complexity: Large Number of Distinct Traces and Activities in event logs

Indeed, educational processes are unstructured and flexible by nature, with a lot of heterogeneous and distinct traces, reflecting the high diversity of behaviors in students' learning paths. Consequently, existing process mining techniques generate highly complex models (called spaghetti models) that are often very confusing and difficult to understand. Moreover, conformance and compliance checking may be complicated with heterogeneous and large scale event logs. One reason for such a result can be attributed to constructing process models from raw traces without due pre-processing. The adoption of filtering, abstraction or clustering techniques may help reducing the complexity of the discovered process models [12], [14], and hence their verification using conformance checking. The issue is to adopt a combination of simplification techniques which reduce the complexity of event logs without losing pertinent information allowing us to discover key concepts and process patterns from these logs.

Concept drift

Usually, when processes are mined and reconstructed from event logs, classic process discover techniques assumed that these processes were stable over the time of observation. But this might not be the case in educational processes. It is not uncommon for the subjacent curriculum to evolve over time and go through major changes from time to time. In fact, courses and study curriculums may be created, modified (e.g., identifier, name, content or structure) or deleted at any time during learning paths of students. Concept drift refers to

TABLE I. EXAMPLE OF AN EDUCATIONAL EVENT LOG STYLES

Matricule	Profil	Training_id	Training label	Training orga_id	Start_date	End_date
7	consultant	tr850	Excel e-learning	Org 13	11/07/11	31/12/11
13	consultant	Tr1923	C++ advanced	Org 135	03/04/12	05/04/12
14	consultant	tr813	Xml basics and XPath	Org 135	04/04/12	06/04/12
…	…	…	…	…	…	…

the situation in which the process changes while being analyzed [16]. An approach to deal with concept drift, in the context of process mining, was introduced in [16].

Interpretation of results by the end users

The models obtained by process mining discovery algorithms have to be comprehensible and useful for the end users' decision-making. For this purpose, visualization techniques and notation simplification are very useful for showing results in a way that is easier to interpret. For instance, instead of showing the whole obtained process model, as directly displayed by process mining algorithms, it is better to abstract its representation using suitable notations (e.g., usual academic notation) understandable by end users or in the form of a list of suggestions, recommendations and conclusions about the obtained results. Also, the interactive interfaces presented to the end-users have to be such as to facilitate the selection of the specific mining method to use with appropriate values for the key parameters to obtain good results/models. Moreover, the trustworthiness of the results should always be clearly indicated [14].

III. ANALYZING EDUCATIONAL PROCESSES USING PROCESS MINING TECHNIQUES

A. Motivating example

Our motivating example is based on real-world training databases from a worldwide consulting company. This company has around 6 000 employees that are free, during their careers, to take different training courses aligned with their profiles. These trainings are provided by internal or external organizations. The data collected for analysis includes the employees' profiles (identifier, function, and number of years of service), their careers (i.e., the jobs/missions they did) and their training paths (the set of training courses taken during the past three years) (see Table I). Training mangers aim to gain more insight in employees' training paths and motivation so they can offer more personalized training courses, according to the job market needs.

In this section, we show how process mining techniques can be used to analyze the training processes underlying this dataset.

B. Preprocessing phase

Data pre-processing allows the transformation of original data into a suitable shape to be used by process mining algorithms. In our case study, the data being collected for analysis is stored in various databases. So, as a first step, we construct a consolidated log (stored as a CSV file) extracted from these databases using an ETL (i.e., *Extract, Transform and Load*) tool, gathering all employees' training courses and work experiences over the last three years. Given that we use the ProM 6.3 framework in process discovery and analyses, we transform this log into the MXML (Mining eXtensible Markup Language) format by using the ProM Import plug-in. Let us note that in our case, during this transformation, we stipule that an employee identifier corresponds to a process instance identifier (an employee training path is understood as a process instance). To obtain a less complex event log, we can use the variety of log filter plug-ins existing in the ProM framework [11]. For instance, the *Event log filter* plug-in enables the selection of only the desired activities in an event log. The *Log filter using simple heuristic* enables a user to select the most frequent activities appearing in an event log.

C. Dotted Chart Analysis

As a first step in our study of the training courses' dataset, we use the dotted chart plug-in of ProM to gain some insight in the underlying process and its performance. The dotted chart shows the spread of events over time by plotting a dot for each event in an event log which enables visually examining an event log and so highlighting some interesting patterns present in it [17]. The dotted chart has two orthogonal dimensions: time and component types. The time is measured along the horizontal axis of the chart. The component types (e.g., instance, originator, task, event type, etc.) are shown along the vertical axis. The dotted chart analysis plug-in of ProM is fully configurable. Based on the chosen component type, the events are rearranged. Fig. 2 illustrates the output of the dotted chart analysis of the training courses' dataset example using process instances as component type. In this chart, every row corresponds to a particular case of the training process, i.e., all the training courses followed by one employee during the last three years. Each training course is represented by two dots of the same color (one per starting date and ending date)

All the instances (one per trainee) are sorted by the first events of trainings, i.e., trainings are sorted by the first date

of their occurrence. We can clearly see from Fig. 2 that each year, there are few trainings scheduling around the last three months (see inside the black circle). Also, almost no training course is scheduled during the summer (see inside the red circles).

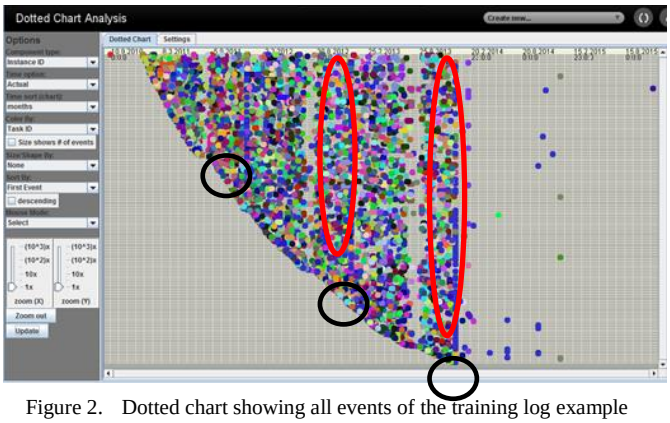


Figure 2. Dotted chart showing all events of the training log example

As a second step, we apply a process model discovery algorithm on only a fragment of our dataset example containing employees' training courses over one year, to get a big picture about the nature of professional training processes. The process model was constructed (using the Heuristic Miner plug-in of ProM [11]) based on an event log containing 8884 events, 2272 training courses performed by 404 different training providers. We can see that the obtained result is an unreadable spaghetti like process model (see Fig. 3).

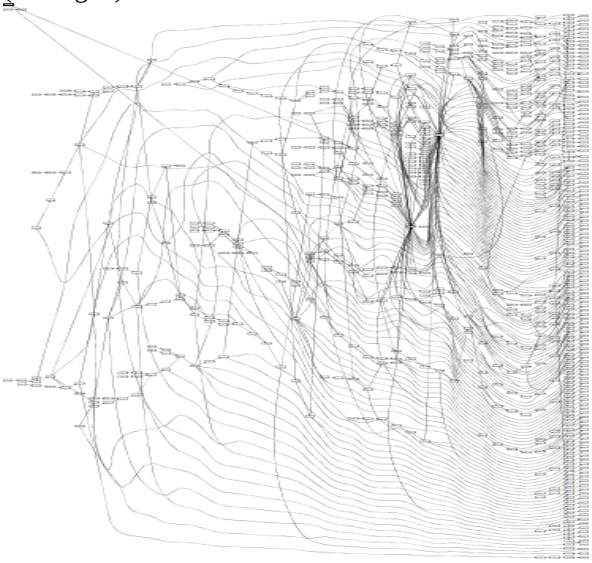


Figure 3. Fragment of a spaghetti process describing all training courses followed by 2980 employees during one year.

D. Social Network Mining

In this section we show the use of the social miner algorithms [10] implemented in the ProM 6.3 framework [11], to examine and assess interactions between training providers and between training courses following their involvement in students' training paths.

According to [18], a social network is defined as a network of interactions or relationships (represented as edges) between entities (represented as nodes). Social Network Analysis (SNA) refers to the collection of methods, techniques and tools in sociometry aiming at the analysis of the structure and composition of ties in social networks. The results of SNA might be used to [18]:

- Identify individuals that are communicating more often with each other (community)
- Identify the individuals with (a) the most outgoing connections (influence), (b) the most incoming connections or in degree (prominence), (c) the least connections (outlier)
- Identify the individuals or groups who play central roles.
- Distinguish bottlenecks (central nodes that provide the only connection between different parts of a network), as well as isolated individuals and groups.

In the EDM context, SNA is usually used to evaluate interactions between students in their collaborative learning tasks, communication actions and online discussions. It can help to understand the group dynamics (structure and content) of educational communities or to quantify the performance of students in teamwork [19-20].

In our case study, we aim to mine and analysis key interaction patterns between training providers and training courses using social mining techniques. To analyze these notions, we rely on two important SNA measures [18]:

Degree Centrality of a node (i.e., the number of nodes that are connected to it): This measure represents the popularity of a node (in our case, training courses or training providers) in a community (in our case, training paths or curriculums).

Betweenness Centrality of a node: In social network context, a node (i.e., training provider or training course) with high betweenness centrality value means that it performs a crucial role in the network, because this node enables the connection between two different groups (i.e., two different training paths or curriculums). If this node is the only bridge linking these two groups and for some reason this node is no longer available, the change of information and knowledge between these two groups would be impossible.

In the process mining field, social mining techniques aim to extract social networks from event logs based on the observed interactions between activities' performers (i.e., resources), depending on how process instances are routed between these performers. These interactions can be generated following one of these five kinds of metrics: (1) transfer of work, (2) delegation or subcontracting of tasks, (3) frequent collaboration (working together) in cases, (4) similarity in executed tasks and (5) reassignment of tasks.

According to our case study, we apply these various metrics to mine social networks between training providers and training courses. Our goal is to find the most pertinent metric allowing us to deduce key interaction patterns between training providers or training courses involved in employees' learning paths. Let us note that, in order to generate social networks between training courses, we replace originator IDs by training IDs of the same events during the event log conversion step in *ProM import*. In what follows, we use the social network plug-ins of ProM 6.3 (based on the four metrics mentioned above) to generate social networks. In the resulting graphs, each node represents a training provider (resp. a training course) where the names have been anonymized for privacy reasons. The oval shape of the nodes in a graph visually expresses the relation between the in and out degree of the connections (arrows) between these nodes. A higher proportion of ingoing arcs lead to more vertical oval shapes while higher proportions of outgoing arcs produce more horizontal oval shapes. We use different views (a ranking view, a stretch by degree ratio, etc.) and two SNA measures (i.e., degree centrality and betweenness centrality) when generating these graphs depending on the key concepts and patterns we want to extract.

Handover of work metric:

Within a case (i.e., process instance) there is a handover of work from individual *i* to individual *j* if there are two subsequent activities where the first is completed by *i* and the second by *j*. In our case, this metric allow us to discover the flow of trainees (specified by the direction of the arrows) between training providers and courses. For instance, in Fig. 4, two providers are connected if one performs a training course causally followed by a course performed by the other provider. In Fig. 4, we distingue two groups of providers strongly related to each other (clustered in cliques) following their causal involvement in training paths. Training providers without arc are those which offer very stand-alone training courses without causal dependency with others. In Fig. 5, we distinguish the most important training courses (trainings with Id 4 et 1) which play central roles in training paths. Training courses or providers with high betweeness centrality represent the ones which connect two different learning paths. In Fig. 6, the size of training courses (with high betweenness) indicates their crucial role as a bridge (i.e., intermediate trainings) between different types of training courses.

Using SNA's measures (betweenness, degree), we can deduce that:

- Training providers or courses with *high degree* are the most popular and prestigious ones, playing a central role in training paths.
- Training providers or courses with no connection with others represent outliers, providing very specific skills, not involved in training paths.
- Nodes with no incoming arcs are training providers (or training courses) who only initiate learning processes (i.e., give the basics for training paths), while nodes with no outgoing arcs are training providers (or courses) who perform only final trainings (i.e., complete training paths with the most required skills).
- Training courses strongly connected to each other hint popular or typical curriculums (or learning paths). The direction of the edges gives the order of training courses followed by students in such curriculums.
- Training courses or providers (with high betweenness) indicates their crucial role as a bridge (i.e., offering intermediate trainings) between different types of training paths.

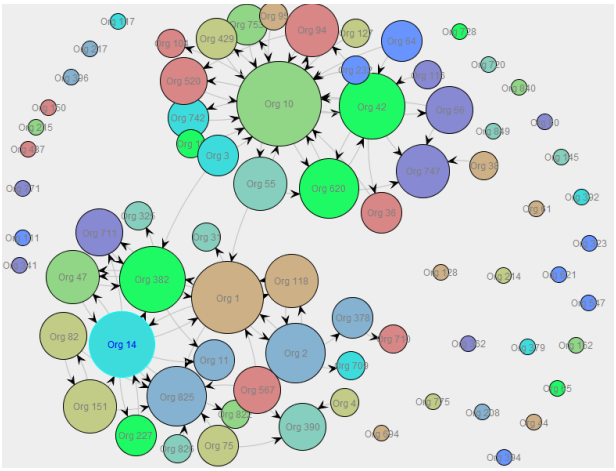


Figure 4. Social network showing handovers between providers of the top 80% of followed training courses using a size by ranking view i.e., the size of a provider's node (which depends on its degree) indicates the importance of its involvement in training paths.

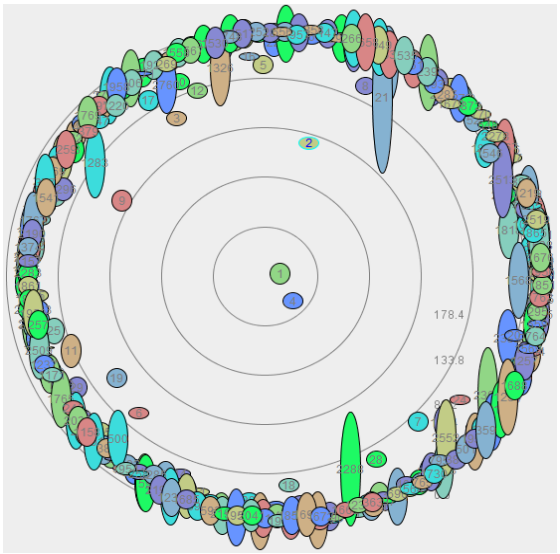


Figure 5. Social network showing handovers between training courses using (1) a ranking view on degree, i.e., courses most involved in training paths are more central in the graph, and (2) a stretch by degree ratio, i.e., the oval shape of the courses' nodes indicates their position in training paths flow.

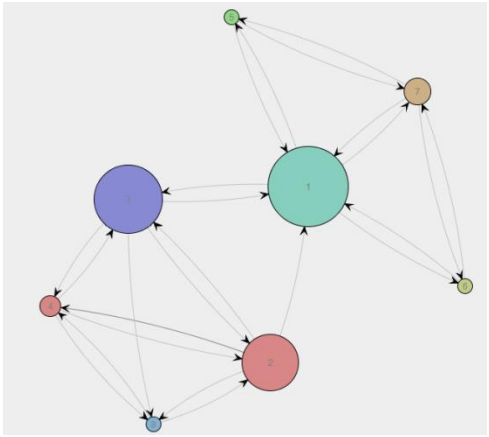


Figure 6. Social network showing handovers between the top 60% of followed training courses, using a ranking on betweenness centrality and a size by ranking view.

Subcontracting metric: A resource *i* subcontracts a resource *j*, when in-between two activities executed by *i* there is an activity executed by *j*. In this case, the start node of an arc represents a contractor and the end node means a subcontractor (see Figs. 7 and 8). In our case study, this metric allow us to extract complementary patterns between training courses and providers.

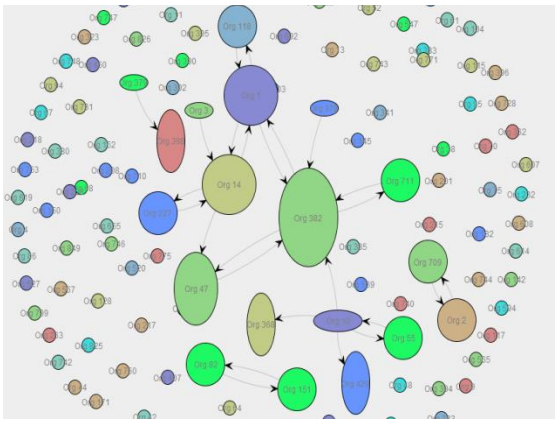


Figure 7. Social network showing subcontracting between training providers of the top 90% of followed training courses.

Using SNA measures, we deduce that:

- Nodes (i.e., training providers or courses) with a high out-degree of centrality (indicated by a horizontal oval shapes) usually play the role of contractors (the main providers or trainings, which give basic skills in these training paths).
- Nodes (i.e., training providers or courses) with a high in-degree of centrality (indicated by a vertical oval shapes) usually act as subcontractors (providers or trainings, which give complementary notions or skills allowing to enhance the notions given by contractors in these training paths).

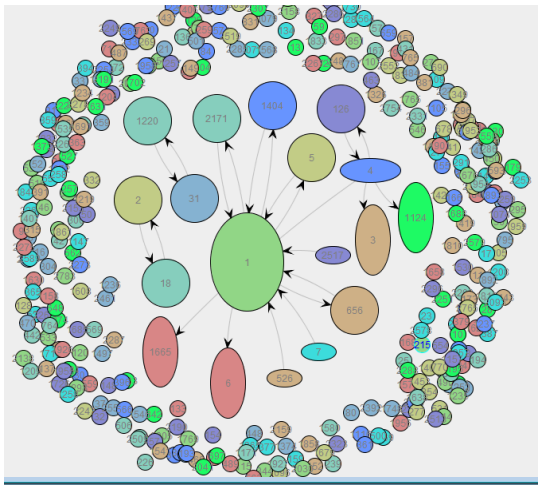


Figure 8. Social network showing subcontracting between the top 80% of followed training courses.

Working together metric: This metric ignores causal dependencies but simply counts how frequently two recourses are performing activities for the same case (see Figs. 9 and 10). In this case, *high density* means that a lot of training providers or courses are involved together in training paths. We can deduce from this social network the most popular curriculums (training providers or courses that work together, i.e., are involved together in training paths). The only difference with the handover metric is that this latter gives us the order followed by students in such curriculums.

Similar task metric: This metric determines who performs the same type of activities in different cases. In our study case, this metric makes sense only to generate relationship between training providers (see Fig. 11). In this case, it allows us to detect training providers who perform the same kind of trainings in curriculums.

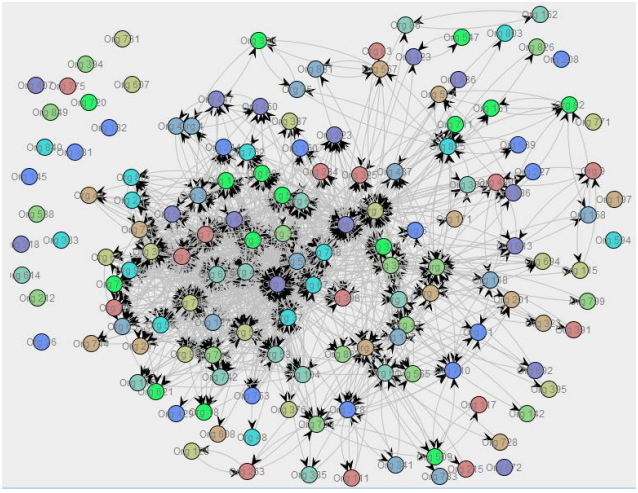


Figure 9. Social network based on working together between training providers. Providers that are often involved together in training paths are related and clustered in cliques. Training providers without arc are those which offer very stand alone rainings.

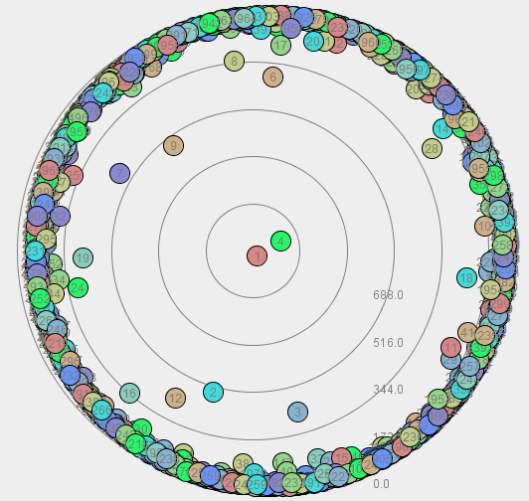


Figure 10. Social network based on working together between training courses using a rancking view on degree, i.e., courses most involved together in training paths are more central in the graph.

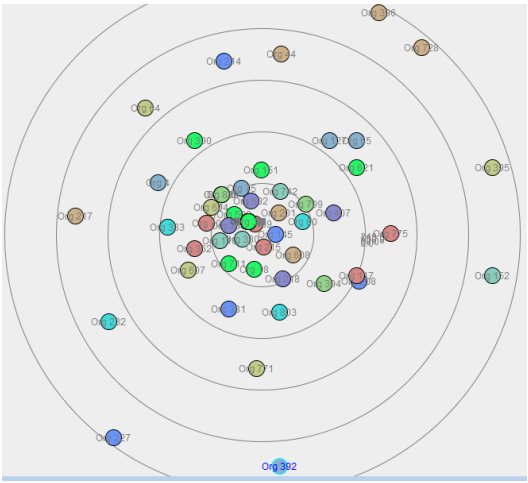


Figure 11. Social network (based on similarity of tasks) between training providers of the top 70 % followed training courses using a ranking view on degree, i.e., providers who perform the most similar collection of trainings are grouped together in the center of the graph.

This experience shows that social network analysis based on event logs is a powerful tool for analyzing coordination patterns between training courses and training providers [9]. Such an approach can also be used to mine interesting patterns about students' behaviors in on-line environments based on resources' usage logs and various interaction logs (e.g., with an intelligent tutoring system).

E. *Process model discovery using a Two-step Clustering Technique*

In order to handle the complexity and heterogeneity of the training paths encountered in the education domain, we propose a two-step clustering approach as a preprocessing step. Our goal is to identify the best training paths by dividing a training event log into homogenous subsets of cases following both their structural similarity and an employability indicator indicating the effectiveness of a training path. In our two-step clustering approach, training paths are firstly partitioned following performance indicators (employability factor and period of unemployment) then training path in each obtained cluster are partitioned further following their structural similarity (see Fig. 12).

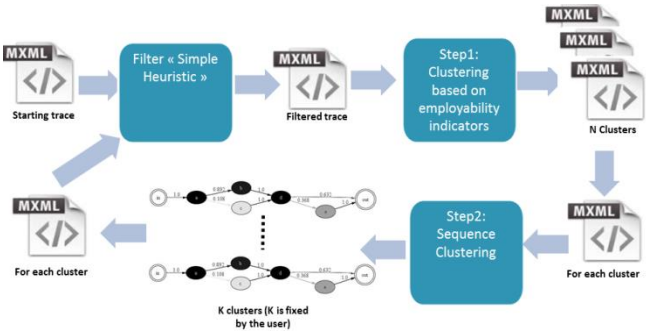


Figure 12. The two-step clustering procedure

A. *First step*

This step consists of creating clusters of similar trainees' profiles based on a training path performance indicator expressed via two criteria. The first one, called employability, concerns the matching between the obtained skills after a training course and those required by a mission. The second criterion represents the time period between a training course followed by an employee and a new mission on which the employee is staffed after it.

a) *Matching criterion:* The criteria that models the matching between skills acquired during a training course and the ones required for a given job/placement, is considered as a real number included between 0 and 1. Hence, this criteria do the matching between a training course followed by an employee, with an identifier « $i \in \{1, \dots, 3340\}$ », and a job/placement will be noted « A_i » with $A_i \in (0,1)$. The set of skills obtained by an employee, identified by i, during his/her trainings is expressed as follows :

$$\mathcal{F}^i = \{F_1^i, F_2^i, \dots, F_{n_i}^i\}$$

Where n_i is an integer greater than or equal to 1. Generally, n_i is at least equal to 3 and less than 10. We note also that for all $\in \{1, \dots, n_i\}$, « F_j^i » indicates that the training course « j » is followed by the employee « i ». For example, $F_1^{10} = Anglais$ means that the employee «10» has followed the English training course. In the same way, the set of skills required by a given job/placement on which the employee « i » has been staffed is noted as follows:

$$\mathcal{M}^i = \{M_1^i, M_2^i, \dots, M_{m_i}^i\}$$

Where m_i is an integer greater than or equal to 1. Generally it is equal to 4 or 5. Also, for all $k \in \{1, \dots, m_i\}$, « M_k^i » indicates that the skill number « k » is required for the job under consideration. For instance, $M_1^{10} = Anglais$ means that the found job/placement for the employee number « 10 » requires English language skill. In addition, the required skills by a given job/placement are weighted according to their importance for the success of this job. This weighting is modeled as follow:

$$\mathcal{P}^i = \{P_1^i, P_2^i, \dots, P_{m_i}^i\}$$

Where for all $j \in \{1, \dots, m_i\}$, $0 < P_j^i < 1$ is the weight associated to the competence « M_j^i » and $\sum_{k=1}^{m_i} P_j^k = 1$. Therefore, the matching criteria between skills obtained by training courses and skills required for a given job/placement is calculated by the following formula:

$$A_i = \sum_{k=1}^{m_i} P_j^i \times \mathbb{I}_{\{M_k^i \in \mathcal{F}^i\}}$$

With $\mathbb{I}_{\{M_k^i \in \mathcal{F}^i\}}$ is an indicator computed by the following rule:

$$\mathbb{I}_{\{M_k^i \in \mathcal{F}^i\}} = \begin{cases} 1 \; si \; M_k^i \in \mathcal{F}^i \\ 0 \; si \; M_k^i \notin \mathcal{F}^i \end{cases}$$

Hence, the distribution characterizing this matching criterion, using our training catalogue and employee information recorded in our example training courses' dataset, is given Fig. 13.

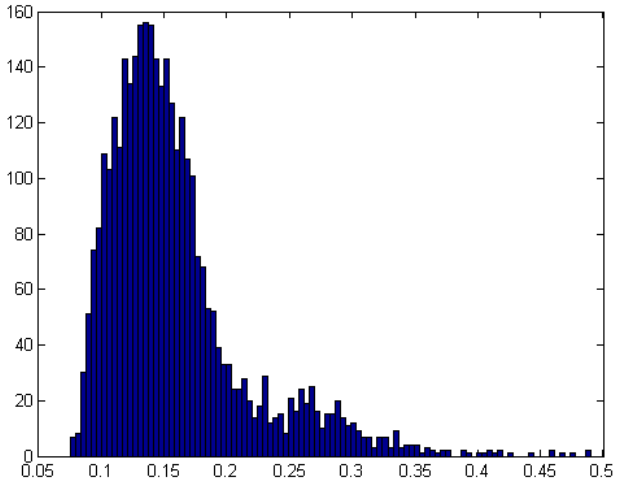


Figure 13. Matching ditribution between training courses and jobs for the employees of our training dataset example.

b) *Time period between training courses and jobs/placements:* This criterion represents, for an employee *i*, the time period between the end of a given training course and the start of his/her next job/placement. This criterion follows the *log normale* probability law. This law is widely used in the modeling of survivor duration. In fact, using the durations, expressed in working days, we obtain the estimated parameters for the used log normale law as follow:

$$\hat{\mu} = 3.16445 \;\; [3.11872, 3.21018]$$
$$\hat{\sigma} = 1.12863 \;\; [1.09721, 1.16191]$$

The graphic representation of the fit of this law is given in Fig. 14. Let us note that we normalize the durations according to the max one in order to have a criterion value comprised between 0 and 1. The goal of this normalization is to homogenize the duration criterion with the matching one.

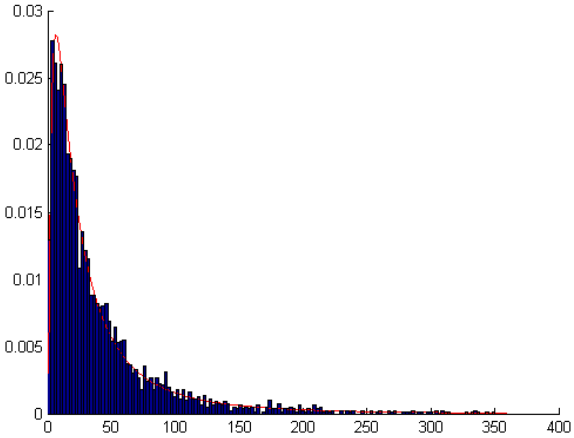


Figure 14. Density probability of the log normal law describing the time between training courses' end and the beginning of new jobs for the employees of the training courses' dataset example

c) *Clustering according to duration and matching criterion:* In these experiments, we do clustering based on the matching and duration criteria defined below. This clustering will help us identify class of training paths for employees that allow them to be staffed on jobs shortly after a training course.

Definition of the cluster number: To get these classes we use the « K-means » technique [21-22], where the optimal number of clusters is determined using a method based on the average silhouette of many clustering where the number of the clusters is varied (the number of clusters *K* is varied between 2 and 5). For more details on this silhouette method, interested readers may refer to [23]. The obtained results are presented in the Fig. 15. When analyzing this figure, we identify a breaking down

of the progression of the average silhouette when *K=3*, this means that the clustering is optimal when we do a clustering with 3 partitions (i.e., clusters).

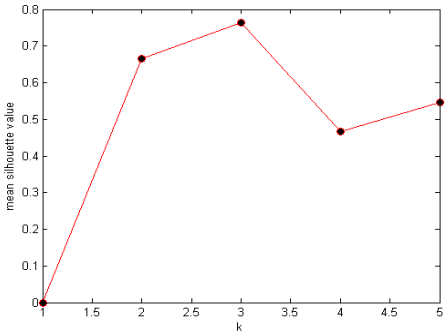


Figure 15. Silhouette Graphical analysis is used to determines the optimal number of clusters. X-axis represents number of clusters and Y-axis indicats asscoiated Silhouette scores.

Clustering for K=3: According to the results obtained in the previous analysis, we apply the "*K-Means*" method, based on the matching and duration criteria, with *K=3*, on our training courses' dataset example. The obtained results are given in Fig. 16.

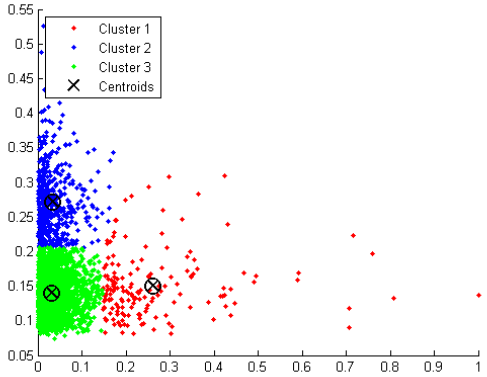


Figure 16. Results of the K-means clustering method applied on our training courses' dataset example using the matching and duration criteria. X-axis represents time period (normalized) between training and the next job. Y-axis corresponds to employability score in (0,1).

K-means method combined with Silhouette Graphical Analysis show that we can identify three trainees groups. Cluster number 2 (Blue group in Fig. 16) represents efficient trainees with high employability score and small employability duration. Instead of, Cluster 1 (red points in Fig. 16) corresponds to inefficient trainees who have small employability score and need more time to find a new job. Finally, green cloud points in Fig. 16 exhibits Cluster number 3 which regroups medium trainees who find quickly a new while they have a small employability score.

We use the fuzzy miner plug-in of ProM (given its robustness to noises) to discover the process model from the training traces of the trainees grouped in the first cluster. We obtain clearly identifiable training paths, as illustrated in Fig. 17. Let us note that these training paths correspond to the least performing ones regarding employability factor and period of unemployment.

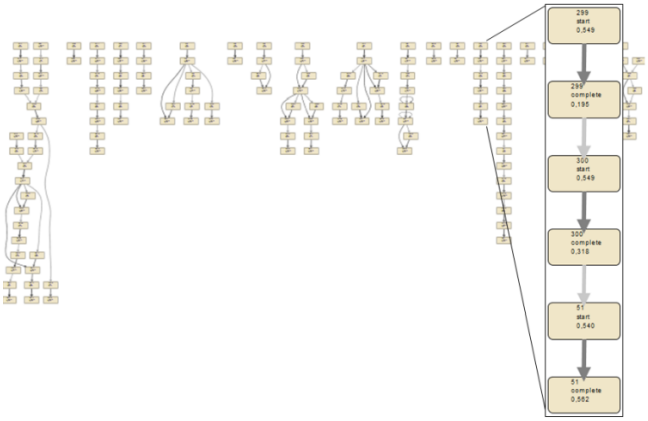


Figure 17. A fragment of the process model showing all the training patterns of cluster 1

Fig. 18 illustrates the process model discovered from the training traces grouped in the second cluster (using the fuzzy miner). Clearly, it is a spaghetti process. The process model discovered from Cluster 3 is even more complex. We can see then that training paths underlying the clusters 2 and 3 are less regular and are more complex that the ones discovered from the first cluster. Let us note that these training paths correspond to the highest performing ones regarding employability factor and period of unemployment.

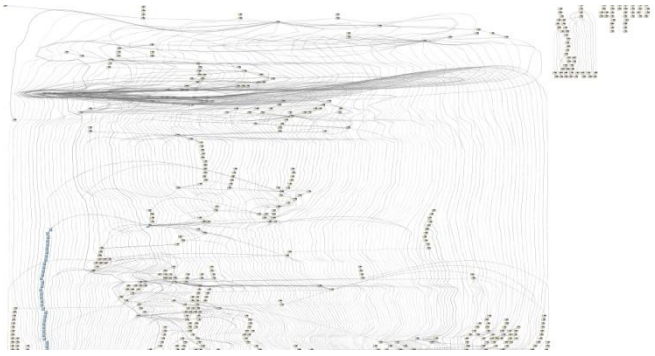


Figure 18. Fragement of the process model (spaghetti-like) underlying cluster 2

In order to obtain simpler training process models, the clusters two and three will be analyzed separately in the second step of our approach. The second step of our approach, simplify further the discovered process models, in the first step, by grouping training paths from each clusters following their structural similarity. Let us note that the first step facilitates the detection of the process patterns and enhances analysis performance because it reduces the searching scope from the whole trainees' information to limit ones for each inferred clusters.

B. Second step

We group training paths (traces in log events) from each of the last two complex clusters discovered in the first step, following their structural similarity using the Sequence clustering technique proposed in [24]. Instead of extracting features from traces, sequence clustering focuses on the sequential behavior of traces. Also, each cluster is based on a probabilistic model, namely a first-order Markov chain. The sequence clustering technique is known to generate simpler models than trace clustering techniques developed in [25]. In our example, when we apply the sequence clustering technique on the second group of trainees with an average employability (i.e., the second cluster of the first step), we obtain three more clusters (cluster 2.1, cluster 2.2 and cluster 2.3). Fig. 19 shows the training process models obtained from the three clusters obtained above, where only transitions occurring above the threshold of 0.05 are represented.

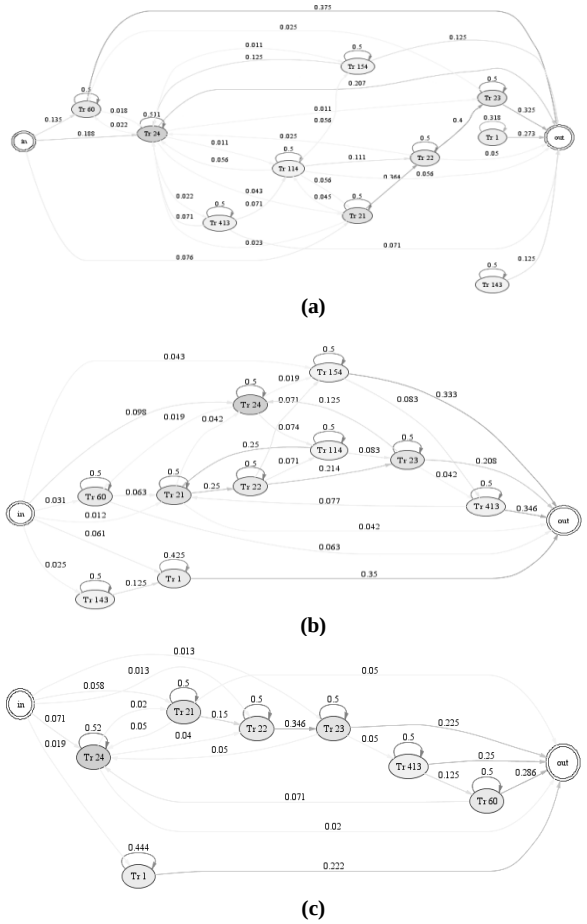


Figure 19. Training process models obtained, from the second cluster of the first step, using the sequence clustering techniques (second step of our approach)

When we apply the sequence clustering technique on the third group of trainees with the less important employability factor (i.e., the third cluster of the first step), we obtain five more clusters (cluster 3.1, cluster 3.2, cluster 3.3, cluster 3.4, cluster 3.5). Fig. 20 shows the training process model underlying the cluster 3.4, where only transitions occurring above the threshold of 0.05 are represented.

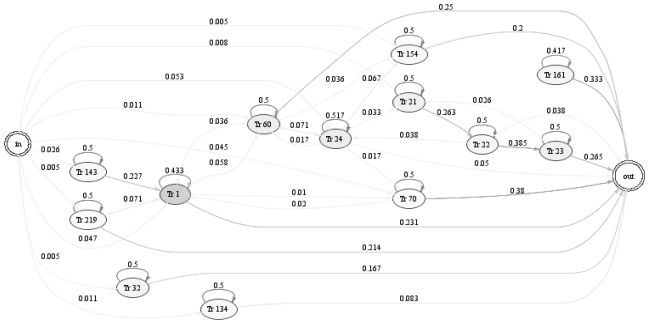


Figure 20. One of the training process model obtained, from the third cluster of the first step, using the sequence clustering techniques (second step of our approach)

IV. PHIDIAS: A PLATFORM FOR DISTRIBUTED EDUCATIONAL PROCESS MINING

To implement our approach, we aim to develop an interactive platform tailored for educational process reconstruction and analysis. This platform will allow different education centers and institutions to load their data and access to advanced data mining and process mining services. Such a platform has to address several issues related to:

- The heterogeneity of the applications and the data sources;
- The connection to some web portals and desktop applications to allow users dealing with the data and exploiting analysis results;
- The ability to add easily new data sources and analysis services;
- The possibility to distribute heavy analysis computations on many processing nodes in order to optimize and enhance platform response time.

To reach these targets we adopt an SOA architecture using an Enterprise Services Bus (ESB) depicted in Fig. 21. This architecture is composed of the following elements: data sources, Enterprise Service Bus, business applications and tools, web services, web portals and connectors. The core of this architecture is the application bus, which guarantees the interoperability and integration of the data sources and applications. For reasons of succinctness we limit the description of the platform to its main components as follow:

Enterprise service bus (ESB)

We have chosen to use ESB architecture in order to have a flexible architecture allowing easily plugging of new applications, data sources and web portals. This integration is done using connectors defining how the data source or the

application will be connected to the bus. We recall that process mining is an application that needs a large number of computations in addition to the required capacities to handle data integration and run business and web server applications. This is why we add a resources optimization layer to the ESB. At this level we consider many features in addition to memory and CPU like transmission times.

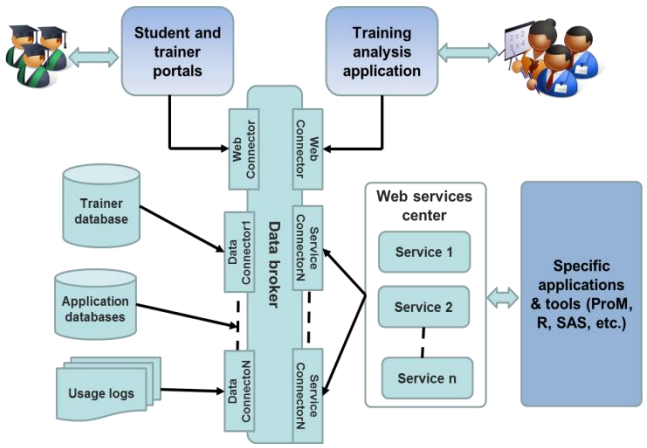


Figure 21. PHIDIAS Architecture

Connectors

The connectors are software components which define how each part of the system will communicate with the ESB. It is one of the main parts of the architecture allowing the system to be flexible and extensible. Its role is the intermediation between a data source, software components, and the web services and applications; and the ESB. This intermediation has different forms; this is why we define three kinds of connectors as follow:

- ***Data connectors:*** these connectors define how transforming the data from their original format to the one for the ESB and how to execute queries on the original sources and then return the results to the ESB.
- ***Web connectors:*** to connect a web portal to the ESB we need to manage the transformation of the user actions recuperated from the HTML blocs to some actions executed on the ESB. This transformation consists of converting for example a user click on a web page to a call of some functions executed on the ESB, then the results are returned back and the connector has to format them in HTML tags.
- ***Service connectors:*** for each service we associate a connector. This connector will store the service interface and transform all the ESB calls to the correct call with the correct parameters. Then it recuperates the results and transforms them to the standard format of the ESB used.

V. RELATED WORKS

Clustering techniques can be used as a preprocessing step to handle large and heterogeneous event logs by dividing an event log into homogenous subsets of cases following their similarity [25-26], [14-27]. One can then discover simpler process models for each cluster. For this purpose, several clustering techniques have been developed and implemented in ProM, such as the Disjunctive Workflow Schema (DWS) plug-in [26] and the Trace Clustering plug-in [28]. Moreover, in [27], the authors propose a combination of trace clustering and text mining to enhance process discovery techniques such that: (1) Trace clustering is applied with the purpose of dividing the event log traces into sub-event logs; (2) A combination of text mining and data mining is proposed with the purpose of finding interesting patterns for the atypical cases. In [29], the authors propose an approach that uses the starting time of each process instance as an additional feature to those considered in traditional Clustering in Process Mining approaches. By combining control-flow features with the starting time, the clusters formed share both a structural similarity and a temporal proximity. Sequence clustering technique was proposed in [24]. This technique differs from Trace Clustering in several ways. Instead of extracting features from traces, sequence clustering focuses on the sequential behavior of traces. Also, each cluster is based on a probabilistic model, namely a first-order Markov chain. Despite of its success, clustering of event logs still remains a subjective technique. A desired goal would be to introduce some objectivity in partitioning the log into homogenous cases. We found out that the sequence clustering technique seems to be the most adequate to partition efficiently training event logs, in the second step of our approach. However, there are some questions we have to investigate when partitioning the process mining problem into smaller problems such as how to combine the results of the individual sub-problems into solutions for the original problems [14]. An important point to discuss when using decomposed process discovery, is how to assess the quality of a decomposition before starting the time-consuming actual discovery algorithm. In [14], the authors defined three quality notions (cohesion, coupling and balance) that can be used to assess a decomposition, before using it to discover a model or check conformance with.

Recently, Educational Process mining or Curriculum mining has emerged as a promising and active research field in Educational Data Mining, dedicated to extracting process related-knowledge from educational datasets. Beyond limitations of EDM, EPM enables greater insights into underlying educational processes. For instance, in Pecheniskiy et al [30], process mining tools, such as process discovery and analysis techniques, were used to investigate the students' behavior during online multiple choice examinations. In [31], the authors use process mining techniques to analyze a collaborative writing process and how the process correlates to the quality and semantic features of the produced document. Analysis techniques were also applied to check the conformance of a set of predefined constraints (e.g., prerequisites) with the event logs. In [6], the authors proposed a technique relying on a set of predefined pattern templates to extract pattern-driven

education models from students' examination traces (i.e., by searching for local patterns and their further assembling into a global model). Under the project "CurriM" [6-7], the authors developed the first software prototype for academic curriculum mining, built on the ProM framework. This tool monitors the flow of curriculums in real-time and return warnings to students (before taking new courses) if prerequisites are not satisfied. Recently, two clustering approaches were proposed in [20], grouping students relying on their obtained marks and their interaction with the Moodle's course. Their aim was to improve both the performance and readability of the mined students' behavior models in the context of e-learning. Finally, in our previous work [9], to handle traces heterogeneity issue, we showed how by associating semantic annotations to educational event logs, we can bring educational processes discovery to the conceptual level. In this way, more accurate and compact educational processes can be mined and analyzed at different levels of abstraction.

VI. CONCLUSION

In this paper, we studied the potential of process mining techniques in the educational domain. Particularly, we show how social mining techniques (implemented in ProM 6.3) can be used to examine and assess interactions between originators (training providers), training courses or pedagogical resources, involved in students' training paths. We also proposed a two-step clustering approach to extract the best training paths depending on an employability indicator. Our future work will continue in several directions. We intend to combine the approach proposed in this paper with other process mining techniques, which allow discovering interaction patterns from email datasets [32] in order to discover interactions patterns between students in their collaborative learning tasks, communication actions and online discussions. Moreover, the proposed architecture will be implemented and deployed and tested on a distributed environment connected to several data sources and applications. This will allow us calibrating and ameliorating our optimization technique either for storage or computation capacities. To enhance the usability of our platform, we are also working on designing an intuitive graphical interface for non-experts that automatically sets parameters and suggests suitable types of analysis. We also plan to conduct a case study that would illustrate the feasibility of process mining approaches in an on-line education setting. Another important step in our works is to investigate further clustering techniques in event logs decomposition to extract typical or atypical training paths depending on domain specific performance indicators and/or on a set of predefined patterns (describing training path templates). We intend also to develop new clustering and classification techniques taking into account semantic annotations on event logs. For instance, trace clustering techniques can be extended to partition event logs depending on trace similarities at the conceptual level. We also intend develop classification techniques to split semantically annotated event logs based on traces' distance from a set of process models or templates, defined at the conceptual level.

ACKNOWLEDGMENT

This work is done by Altran Research and Altran Institut in the context of the project PERICLES (http://e-pericles.org/).

REFERENCES

[1] A. Hicheur Cairns, B. Gueni, M. Fhima, A. Cairns, S. David and N. Khelifa, "Custom-designed professional training contents and curriculums through educational process mining," The Fourth International Conference on Advances in Information Mining and Management (IMMM 2014), Jul. 2014, pp. 53-58.

[2] C. Romero, S. Ventura and E. Garcia, "Data mining in course management systems: moodle case study and tutorial," Computers & Education, 51(1), 2008, pp. 368-384.

[3] C. Romero, S. Ventura, M. Pechenizkiy and R. Baker, "Handbook of Educational Data Mining," Boca Raton, FL: CRC Press, Taylor&Francis, 2010.

[4] T. Calders and M. Pechenizkiy, "Introduction to the special section on educational data mining," ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD), 2011, pp. 3-6.

[5] C. Romero and S. Ventura, "Data mining in education," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol 3, 2013, pp. 12–27.

[6] N. Trčka and M. Pechenizkiy, "From local patterns to global models: towards domain driven educational process mining," Ninth International Conference on Intelligent Systems Design and Applications (ISDA 2009), Nov. 2009, Pisa, Italy, pp. 1114-1119.

[7] N. Trčka, M. Pechenizkiy and W.P.M. van der Aalst "Process mining from educational data (Chapter 9)," In Handbook of Educational Data Mining, CRC Press, pp. 123-142.

[8] W. M. P. Van der Aalst et al., "Process mining manifesto," International Conference on Business Process Management Workshops (BPM 2011), 2011, pp. 169–194.

[9] A. Hicheur-Cairns et al., "Using semantic lifting for improving educational process models discovery and analysis," 4th International Symposium on Data-driven Process Discovery and Analysis (SIMPDA 2014), Italy, Nov. 2014, pp. 150-161.

[10] W.M.P. van der Aalst and M. Song, "Mining social networks: Uncovering interaction patterns in business processes," International Conference on Business Process Management (BPM 2004), Lecture Notes in Computer Science, vol. 3080, Springer, Berlin, 2004, pp. 244-260.

[11] B. F. van Dongen, A. K. A. de Medeiros, H. M. W. Verbeek, A. J. M. M. Weijters and W. M. P. van der Aalst, "The prom framework: a new era in process mining tool support," International Conference on Applications and Theory of Petri Nets (ICATPN'05), Berlin, Heidelberg, 2005, pp. 444-454.

[12] M. Reichert, "Visualizing large business process models: challenges, techniques, applications," 1st International Workshop on Theory and Applications of Process Visualization Presented at the BPM 2012 (TAProViz'12), Tallin, Sep. 2012, pp. 725-736.

[13] J. C. Bose, R. S. Mans and W. M. P. van der Aalst, "Wanna improve process mining results?," IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2013, Singapore, Apr 2013, pp. 127-134.

[14] B.F.A Hompes, H.M.W. Verbeek and W.M.P. van der Aalst, "Finding suitable activity clusters for decomposed process discovery," the 4th International Symposium on Data-driven Process Discovery and Analysis (SIMPDA 2014), CEUR

Workshop Proceedings 1293, Milan, Italy, November 19-21, 2014, pp. 16-30.

[15] J. Munoz-Gama, J. Carmona and W.M.P. van der Aalst, "Conformance checking in the large: partitioning and topology," International Conference on Business Process Management (BPM 2013), 2013, pp. 130-145.

[16] R.P.J.C. Bose, W.M.P. van der Aalst, I. Zliobaite and M. Pechenizkiy, "Handling concept drift in process mining," 23rd International Conference on Information Systems Engineering (CAiSE'2011), Springer, 2011, pp. 391-405.

[17] M. Song and W.M.P. van der Aalst, "Supporting process mining by showing events at a glance," Seventeenth Annual Workshop on Information Technologies and Systems (WITS'07), Montreal, Canada, December 8-9, 2007, pp.139–145.

[18] C. Aggarwal, "Introduction to social network data analytics," In Social Network Data Analytics, Springer, 2011, pp. 1-15.

[19] P. Crespo and C. Antunes, "Social networks analysis for quantifying students' performance in Teamwork," Educational Data Mining (EDM 2012), 2012, pp. 234-235.

[20] G. Obadi, P. Drázdilová, J. Martinovic, K. Slaninová and V. Snásel, "using spectral clustering for finding students' patterns of behavior in social networks," International Workshop on Databases, Texts, Specifications, and Objects (DATESO), vol 567, 2010, pp. 118-130.

[21] G. A. F. Seber, "Multivariate observations," Hoboken, NJ: John Wiley & Sons, Inc., 1984.

[22] H. Späth, "Cluster dissection and analysis: theory, FORTRAN programs, examples," Translated by J. Goldschmidt, New York: Halsted Press, 1985.

[23] L. Kaufman and P. J. Rousseeuw, "Finding groups in data: an introduction to cluster analysis," Hoboken, NJ: John Wiley & Sons, Inc., 1990.

[24] G.M. Veiga and D.R. Ferreira, "Understanding spaghetti models with sequence clustering for ProM," Business Process Management Workshops, vol. 43, 2010, pp. 92–103.

[25] M.Song, C.W. Günther and W.M.P. van der Aalst, "Trace clustering in process mining," Business Project Management (BPM 2008), vol. 17, Springer, Heidelberg, 2009, pp. 109–120.

[26] A.K.A. De Medeiros, A. Guzzo, G. Greco, W.M.P. van der Aalst, A.J.M.M. Weijters, B.F. van Dongen and D. Saccà, "Process mining based on clustering: A quest for precision," Business Process Management International Workshops (BPM 2007), Brisbane, Australia, Berlin: Springer-Verlag, pp. 17-29.

[27] J. Weerdt, J. Vanthienen and B. Baesens, "Leveraging process discovery with trace clustering and text mining for intelligent analysis of incident management processes," IEEE Congress on Evolutionary Computation (CEC), 2012, pp. 1-8.

[28] R.P. Jagadeesh Chandra Bose and W.M.P. van der Aalst, "Context aware trace clustering: towards improving process mining results," In Proceedings of the SIAM International Conference on Data Mining, SDM, May, 2009, pp. 401–412.

[29] D. Luengo and M. Sepúlveda, "Applying clustering in process mining to find different versions of a business process that changes over time," Business Process Management Workshops, 2011, pp. 153-158.

[30] M. Pechenizkiy, N. Trčka, E. Vasilyeva, W.P.M. van der Aalst and P. De Bra, "Process mining online assessment data," Educationnal Data Mining (EDM'09), 2009, pp. 279-288.

[31] V. Southavilay, K. Yacef and R. A. Calvo, "Process mining to support students' collaborative writing," Educational Data Mining conference proceedings, 2010, pp. 257-266.

[32] W.M.P. van der Aalst and A. Nikolov, "EMailAnalyzer: an e-mail mining plug-in for the ProM framework," International Conference on Business Process Management, 2007.

www.iariajournals.org

International Journal On Advances in Intelligent Systems
issn: 1942-2679

International Journal On Advances in Internet Technology
issn: 1942-2652

International Journal On Advances in Life Sciences
issn: 1942-2660

International Journal On Advances in Networks and Services
issn: 1942-2644

International Journal On Advances in Security
issn: 1942-2636

International Journal On Advances in Software
issn: 1942-2628

International Journal On Advances in Systems and Measurements
issn: 1942-261x

International Journal On Advances in Telecommunications
issn: 1942-2601